

MINING FOR KNOWLEDGE TO BUILD DECISION SUPPORT SYSTEM
FOR DIAGNOSIS AND TREATMENT OF TINNITUS

by

Pamela Liberty McDermon Thompson

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Computing and Information Systems

Charlotte

2011

Approved by:

Dr. Zbigniew W. Ras

Dr. Tiffany Barnes

Dr. Mirsad Hadzikadic

Dr. Pawel Jastreboff

Dr. James Studnicki

© 2011
Pamela Liberty McDermon Thompson
ALL RIGHTS RESERVED

ABSTRACT

PAMELA LIBERTY MCDERMON THOMPSON. Mining the tinnitus database for knowledge: design foundations of a decision support system for improving treatment effectiveness based on new feature discovery and action rules. (Under the direction of Dr. ZBIGNIEW W. RAS)

Tinnitus problems affect a significant portion of the population and are difficult to treat. Treatment processes are plentiful, yet not completely understood. In this dissertation, we present a knowledge discovery approach which can be used to build a decision support system for supporting tinnitus treatment. Our approach is based on a significant enlargement of the initial tinnitus database by adding many new tables containing new temporal features related to tinnitus evaluation and treatment outcome. Research presented in this thesis includes knowledge discovery with temporal, text, and quantitative data from a patient dataset of 3013 visits representing 758 unique patient tuples. Additionally, a new rule generating technique and clustering methods are presented and used to develop additional new temporal features and knowledge in this complex domain. Of particular interest is the role that emotions play in treatment success for tinnitus following the TRT method developed by Dr. Pawel Jastreboff. The ultimate goal of understanding the relationships among the treatment factors and measurements in order to better understand tinnitus treatment will result in the design foundations of a decision support system to aid in tinnitus treatment effectiveness.

ACKNOWLEDGEMENTS

I am extremely grateful for the encouragement and support that I have received from the members of my committee and fellow graduate students at the University of North Carolina at Charlotte.

I would like to thank my dissertation supervisor, Zbigniew W. Ras, for his patience, guidance, and for exposing me to his internationally recognized research team, including the founder of Tinnitus Retraining Therapy, Dr. Pawel Jastreboff, and the developer of LISp-Miner, Dr. Jan Rauch.

I would also like to thank the members of my committee (Dr. Barnes, Dr. Hadzikadic, Dr. Jastreboff, and Dr. Studnicki) for their valuable feedback. Many thanks are also due to fellow graduate students Cynthia Zhang and Rory Lewis for sharing their knowledge and providing their support through the years. My thanks are also extended to Dora Bradley; her encouragement and positive outlook have been very much appreciated.

Finally, I would like to thank my two daughters, Jacquelyn and Virginia, for their help at home throughout my research and my deceased mother and father, Virginia Liberty McDermon and Ernest Malone McDermon Jr., for instilling in me a love of learning and a positive outlook on life.

TABLE OF CONTENTS`

LIST OF TABLES	vii
LIST OF FIGURES	ix
CHAPTER 1: INTRODUCTION	1
1.1 Background	1
1.2 Motivation and Approach	3
1.3 Contributions of this Dissertation	5
1.4 Organization of this Document	6
CHAPTER 2: METHODOLOGY	8
2.1 Domain Knowledge	8
2.2 Data Collection	10
2.3 Database Features	13
2.4 Extraction, Transformation and Loading of Original Features	17
CHAPTER 3: NEW FEATURE CONSTRUCTION FOR THE TINNITUS DATABASE	20
3.1 Text Extraction and Mining	20
3.2 Temporal Feature Development and Extraction	22
3.3 Feature Development for Categorical Data	24
3.4 New Features based on the Tinnitus Functional Index and Emotions	24
CHAPTER 4: CLUSTERING TECHNIQUES FOR FEATURE EXTRACTION	27
4.1 Data Selection, Grouping, and Temporal Feature Extraction	28
4.2 Temporal Feature Extraction with Clustered Data	30

4.3 New Temporal Features for Clustered Visits: Coefficients and Angles	33
4.4 Quadratic Equation Based New Features	36
CHAPTER 5: MINING UNCLUSTERED DATA	38
5.1 Original Experiment and Results	38
5.2 Structure of the Decision Attribute	39
CHAPTER 6: MINING CLUSTERED DATA	51
CHAPTER 7: ACTION RULES	56
7.1 Action Rules – Preliminary Research	56
7.2 LISp-Miner for Action Rule Discovery	65
7.3 New Application for Action Rule Discovery	69
7.4 The Tinnitus Functional Index and Emotions	71
7.5 Emotions Feature Development	72
7.6 Data Preparation for Action Rule Discovery	72
CHAPTER 8: ACTION RULES – EXPERIMENT AND RESULTS	75
8.1 Action Rules Ac4ft-Miner with LISp-Miner	75
8.2 Analytical Questions and Rules from LISp-Miner	80
8.3 Action Rules and MARDs	90
8.4 A Comparison of Mining Applications	95
CHAPTER 9: CONCLUSION AND DISCUSSION	96
REFERENCES	99
APPENDIX A: ATTRIBUTES, FEATURES, AND DESCRIPTIONS	102
APPENDIX B: TEST DATA, MARDs	105

LIST OF TABLES

TABLE 1:	New Boolean Features	21
TABLE 2:	Tinnitus Functional Index (Scale of 0 to 10)	26
TABLE 3:	An Example of Calculating Visit Duration	28
TABLE 4:	Categories for Total Score Discretization	40
TABLE 5:	WEKA Results, Classifier Tree for J4.8	42
TABLE 6:	Original Data with Standard Deviations and Averages	44
TABLE 7:	Original Data with Standard Deviations, Averages and Sound	45
TABLE 8:	Original Data with Standard Deviations, Averages and Text	46
TABLE 9:	Original Data with Standard Deviations, Averages, Sound, Text	47
TABLE 10:	Original Data with Text	48
TABLE 11:	Original Data with Sound and Recovery Rate	49
TABLE 12:	Original Data with Sound, Text and Recovery Rate	49
TABLE 13:	Attributes and Features for the Clustered Databases	52
TABLE 14:	Seeds generated with Distance ≤ 3 Weeks	59
TABLE 15:	Seeds generated with Distance ≤ 4 Weeks	60
TABLE 16:	Attributes and Features used in LISp-Miner and Arc4ft-Miner	76
TABLE 17:	Mining Tasks of Interest	77
TABLE 18:	Attribute categories frequency analysis for feature E1	78
TABLE 19:	Hypothesis for Resulting Rule	79
TABLE 20:	TASK 01	81
TABLE 21:	TASK 02	82
TABLE 22:	TASK 03	84

TABLE 23:	TASK 04	85
TABLE 24:	TASK 05	87
TABLE 25:	Categories for Question 21	88
TABLE 26:	TASK 06	89
TABLE 27:	MARDs Input File and Array Loading	94

LIST OF FIGURES

FIGURE 1:	Development of a Vicious Cycle	10
FIGURE 2:	Patient Categories	11
FIGURE 3:	Original Database Description (from Access)	13
FIGURE 4:	New Tinnitus Functional Index Table (from Access)	14
FIGURE 5:	Sound Level Centroid	22
FIGURE 6:	Sound Level Spread	23
FIGURE 7:	Recovery Rate	23
FIGURE 8:	Sample of TFI Question	25
FIGURE 9:	System Overview	27
FIGURE 10:	Matching for Closest Visit Pattern	29
FIGURE 11:	Angle Formulation	35
FIGURE 12:	Points that Make Up the Quadratic Equation	37
FIGURE 13:	Top Classification Results: J48 with Decision Variable TSa and Sound Level Centroid, Sound Level Spread, and Recovery Rate	43
FIGURE 14:	Graph of Table 6	44
FIGURE 15:	Graph of Table 7	45
FIGURE 16:	Graph of Table 8	46
FIGURE 17:	Graph of Table 9	47
FIGURE 18:	Graph of Table 10	48
FIGURE 19:	Graph of Table 12	50
FIGURE 20:	Weka Results	54
FIGURE 21:	Comparison Between Decision Variables	55
FIGURE 22:	Example of an Action Rule	57

FIGURE 23: GUHA Procedure	66
FIGURE 24: Ac-4ft Quantifier	68
FIGURE 25: Input Screen for MARDs	91
FIGURE 26: System Diagram for MARDs	92
FIGURE 27: MARDs Experiment and Action Rules	93

CHAPTER 1: INTRODUCTION

Tinnitus, sometimes called “ringing in the ears”, affects a significant portion of the population. Some estimates show the portion of the population in the United States affected by tinnitus to be 40 million, with approximately 10 million of these considering their problem significant [1]. Many definitions exist for tinnitus. One definition of tinnitus relevant to this research is “. . . the perception of sound that results exclusively from activity within the nervous system without any corresponding mechanical, vibratory activity within the cochlea, and not related to external stimulation of any kind” [2]. Hyperacusis or decreased sound tolerance frequently accompanies tinnitus and can include symptoms of misophonia (strong dislike of sound) or phonophobia (fear of sound). Physiological causes of tinnitus can be difficult or impossible to determine, and treatment approaches vary.

1.1 Background

Tinnitus Retraining Therapy (TRT), developed by Dr. Jastreboff, is one treatment model with a high rate of success and is based on a neurophysical approach to treatment. TRT “cures” tinnitus by building on its association with many centers throughout the nervous system including the limbic and autonomic systems. The limbic nervous system (emotions) controls fear, thirst, hunger, joy and happiness. It is connected with all sensory systems. The autonomic nervous system controls such functions as breathing, heart rate and hormones. When the emotion linked limbic system becomes involved with tinnitus, symptoms may worsen and affect the autonomic nervous system [3]. TRT combines counseling and sound habituation to successfully treat a majority of patients. Conceptually,

habituation refers to a decreased response to the tinnitus stimulus due to exposure to a different stimulus [4]. Degree of habituation determines treatment success, yet greater understanding of why this success occurs and validation of the TRT technique will be useful [5]. Dr. Jastreboff believes there is a strong connection with improvement in emotions and improvement in tinnitus symptoms; this belief has been supported by a 2002 study by Josef P. Rauschecker using magnetic resonance imaging in patients suffering from tinnitus and exploring the limbic system. [6]

The treatment requires a preliminary medical examination, completion of an Initial Interview Questionnaire for patient categorization, audiological testing, a visit questionnaire referred to as a Tinnitus Handicap Inventory, another visit questionnaire known as the Tinnitus Functional Inventory (available for patients in the second dataset), tracking of instruments, and a follow-up questionnaire.[3] Data from a sample of 555 patients was originally presented in a relational database consisting of eleven tables. Data from a combined sample of 758 patients was obtained and added to the analysis. Patient tuples were related to one to many tuples in other tables based on patient visits through the course of treatment. Tuples with data related to treatments during visits were uniquely identified by patient id and visit number and date, enabling temporal treatment of data. The authors focused on cleansing and analysis of existing data, along with automating the discovery of new and useful features in order to improve classification and understanding of tinnitus diagnosis and improvement. The new dataset includes a new Questionnaire called the “TFI” or Tinnitus Functional Index; this will allow for improved study of treatment effectiveness based on new features that can be tied to emotions.

Both datasets represented many challenges in successful mining and analysis. The database, in each case, was created from manual forms from patient visits and treatments transcribed to a relational database format over a period of years. In order to perform research, it was important to understand the domain knowledge related to the problem from the areas of otology, psychology, and computer science. Based on domain knowledge, a determination had to be made on useful features to the problem, and then data had to be consolidated and cleansed with many discrepancies resolved programmatically. Some similar data was stored in different formats, and contained inconsistencies. Null values were programmatically removed and generally not included in the aggregate table. Some of the medical data appeared in nested arrays which is not a suitable data representation for traditional data mining algorithms.

An additional problem was presented with the task of creating a single table of data for mining purposes. Relationships among multiple tables were based on a patient id that was represented in different formats and a visit date and number related to many of the tuples was not consistent across tables.

1.2 Motivation and Approach

This dissertation explores various approaches for mining the tinnitus datasets in order to develop new and relevant temporal and other features. Additionally, text mining and novel clustering techniques are also used with the ultimate goal of rule extraction based on the knowledge gained from the tinnitus database. The knowledge learned will provide the basis for a decision support system designed to improve treatment efficiency and effectiveness for tinnitus sufferers.

The original database was obtained early in the research – it contained information about 555 unique patients. We cleansed, transformed, and mined this database for knowledge using decision tree analysis. Three new features were developed in the process: sound level centroid, sound level spread, and recovery rate. Additionally, text mining and frequent pattern discovery was used to add new Boolean features for cause of Tinnitus (Noise, Stress) and for Prescription Drug Use (Medical). From the visit sequence and total score, many new features were added representing the coefficients of the polynomial equation that maps to the visit sequence and total score plot, and angles are calculated and stored for various combinations of points on this line. Finally, mining was performed using clustered datasets represented either by three or four patient visits. These datasets have been determined by an algorithm that looks at the length of time between visits and matches like sequences. New decision features were used in the mining; these decision features were based on the discretized Total Score from the Tinnitus Handicap Inventory, a questionnaire regularly completed during patient visits.

The new, extended tinnitus database represents information about 758 patients with information repeated from the original database, along with the addition of visits and a new questionnaire, the Tinnitus Function Index. The extended database was mined for comparison to the original work. New patients in the extended database represented those patients that had completed the Tinnitus Function Index; these patient visits were separated and used for mining and action rule discovery based on all features and treatment success indicators including several new features tied to emotions (based on a mapping of questions to Thayer's Arousal-valence emotion plane and the mood model as described by Grekow and Ras [7]).

1.3 Contributions of this Dissertation

In this dissertation, the following new features based on a patient visit sequence are introduced to the tinnitus database: sound level centroid, sound level spread, and recovery rate. Additionally, three new text based features indicating the cause of tinnitus are introduced: Noise, Stress, and Medical. New features for coefficients and angles related to the plot of the line of visit length and Total Score are calculated and also used in mining and rule discovery. Finally, emotion-based features E1 (Energetic-positive), E2 (Energetic-negative), E3 (Calm-negative), and E4 (Calm-positive) are introduced based on questions from the Tinnitus Functional Index for new patients and the Arousal-valence emotion plane [7].

Types of learning that occurred include classification learning to help with classifying unseen examples, association learning to determine any association among features (largely statistical), clustering to seek groups of examples that belong together in order to realize improvement in classification and rule discovery, and action rule discovery.

Many of the new features show promise in mining for use in evaluating the treatment methods and corresponding treatment success for tinnitus sufferers. Additionally, the emotion based features can be used in the continuation of research related to the new and novel music therapy approaches to tinnitus treatment [8]. Ultimately, the important knowledge gained in this study will be used to extend the research and to build a decision support system that be used by physicians treating tinnitus in order to maximize treatment effectiveness by placing more emphasis on the emotional state of the patient.

1.4 Organization of this Document

The remainder of this dissertation is structured as follows:

Chapter 2: Methodology. The extensive domain knowledge and data collection methods are introduced and discussed along with previous work.

Chapter 3: New Feature Construction for the Tinnitus Database. New features for the tinnitus database are presented and explained, along with method of construction and references to previous work.

Chapter 4: Advanced Clustering Techniques for Temporal Feature Extraction. The clustering algorithm for tinnitus visits for improving the performance of classifiers is presented. Additionally, an analysis of available clustering algorithms is evaluated along with the rationale for using the developed clustering algorithm.

Chapter 5: Unclustered Data: Classification Study (J48, Random Forest, Multilayer Perceptron). The unclustered data represents classification with both the original and combined dataset with new features and discretized total score. Results from mining the original, and combined tinnitus data with new features (not including coefficients and angles) are analyzed and compared for improved results. Classifiers for treatment success based on the discretized total score from the Tinnitus Handicap Inventory are presented. The contribution of new features toward classification of tinnitus patients and treatment success is presented based on the results of the decision tree analysis of the unclustered data for the original and combined tinnitus data.

Chapter 6: Clustered Data: Classification Study (J48, Random Forest, Multilayer Perceptron). The clustered data represents the original dataset only with clustering by seed patient into three and four visit sets with new features including coefficients and angles and

discretized total score. The new clustering algorithm is used with the new features developed for coefficients and angles calculated from the line plotted from visit sequence and Total Score (Tinnitus Handicap Inventory) for three and four visit sets. Classifiers for treatment success based on the discretized total score from the Tinnitus Handicap Inventory with the new coefficients and angles features have been built. The contribution of clustering and new features toward classification of tinnitus patients and treatment success is presented based on the results of the decision tree analysis of the clustered three and four visit data for the original tinnitus data.

Chapter 7: Action Rules. New patient data is separated from the combined database to learn action rules for treatment success based on visits. Of particular interest is the contribution of the new Tinnitus Functional Index and the emotion based features developed from the index.

Chapter 8: Action Rules Experiment and Results. The experiments and results from the action rules study will be presented in this chapter.

Chapter 9: Conclusion and Discussion. A summary of the accomplishments achieved in this research is discussed along with the contribution toward a decision support system for tinnitus. Plans for future research are presented.

CHAPTER 2: METHODOLOGY

2.1 Domain Knowledge

The domain knowledge for tinnitus involves many disciplines, including psychology and otology. Psychiatrists tend to focus on the hallucinatory type of phantom perception associated with tinnitus, while otolaryngologists are interested on the tonal or noise like perceptions associated with the condition.

Tinnitus appears to be caused by a variety of factors including exposure to loud noises, head trauma, disease (diabetes, Lyme disease, others), and muscle tension. An interesting fact is that Tinnitus can be induced in 94% of the population by a few minutes of sound deprivation [9].

Decreased sound tolerance frequently accompanies tinnitus and can include symptoms of misophonia (strong dislike of sound) or phonophobia (fear of sound). Physiological causes of tinnitus can be difficult or impossible to determine, and treatment approaches vary. Past approaches to treatment tend to have been based on definition, and treatment often focused on tinnitus suppression. Suppression is accomplished by using a listening device set to a mixing point to suppress tinnitus. The mixing point is that point where the sound from the listening device masks or suppresses the sound from the tinnitus.

Jastreboff offers an important new definition (hence treatment) for tinnitus that focuses on the subjective aspect of the condition and describes tinnitus as resulting

exclusively from activity within the nervous system that is not related to corresponding activity with the cochlea or external stimulation.

Tinnitus Retraining Therapy (TRT), developed by Jastreboff, is a treatment model with a high rate of success and is based on a neurophysical approach to treatment.

Neurophysiology is a branch of science focusing on the physiological aspect of nervous system function [3]. Tinnitus Retraining Therapy “cures” tinnitus by building on its association with many centers throughout the nervous system including the limbic and autonomic systems.

The limbic nervous system (emotions) controls fear, thirst, hunger, joy and happiness. The limbic nervous system is connected with all sensory systems. The autonomic nervous system controls many functions such as breathing, heart rate and hormones. When the limbic system becomes involved with tinnitus, symptoms may worsen and affect the autonomic nervous system [3]. Unfortunately, many patients seeking treatment other than Tinnitus Retraining Therapy are often told that nothing can be done about their tinnitus. This has the effect of causing a limbic nervous system reaction, which then, over time, can cause strengthening of the negative affect of the tinnitus on the patient (see Figure 1: Development of a Vicious Cycle) [3].

Development of a Vicious Cycle

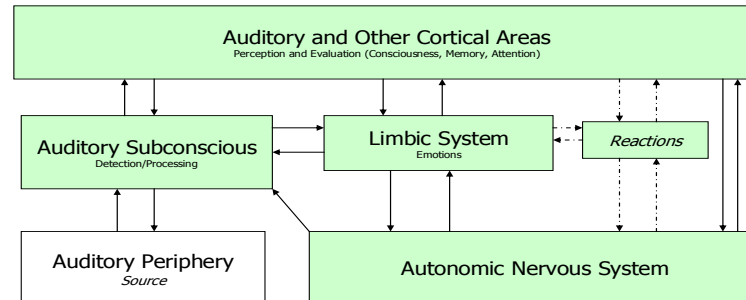


Figure 1: Development of a Vicious Cycle

Tinnitus Retraining Therapy combines medical evaluation, counseling and sound habituation therapy (rather than suppression) to successfully treat a majority of patients. Conceptually, habituation refers to a decreased response to the tinnitus stimulus due to exposure to a different stimulus [10]. The goal of habituation is to reduce the emotional reaction to tinnitus, and eventually eliminate the awareness of tinnitus altogether. Degree of habituation determines treatment success, yet greater understanding of why this success occurs and validation of the Tinnitus Retraining Therapy technique will be useful. The ultimate goal is to lessen or eliminate the impact of tinnitus on the patient's life. It is important to note that Tinnitus Retraining Therapy can often take years to complete.

2.2 Data Collection

A preliminary medical evaluation of patients is required before beginning Tinnitus Retraining Therapy. Data from the medical evaluation is not directly included in the data presented to the researchers. Much of this data contain information subject to privacy concerns, a consideration of all researchers engaged in medical database exploration. Some information, however, is included in comment type features which describe medications the

patient may take and other conditions that might be present, such as diabetes. One feature includes text information on the patient's perceived cause of the onset of tinnitus.

After the medical evaluation, the completion of an Initial Interview Questionnaire for patient categorization is completed. This questionnaire collects data on many aspects of the patient's tinnitus, sound tolerance, and possible hearing loss. The interview also helps determine the relative contribution of hyperacusis, misophonia and phonophobia. Questions relate to activities prevented or affected (concentration, sleep, work, etc.) for tinnitus and sound tolerance, if a hearing aid is worn, levels of severity, annoyance, effect on life, and many others. All responses are included in the database. Audiological testing is performed to determine left and right ear pitch, loudness discomfort levels, and suppressibility along with other measures.

Based on information from the medical evaluation and the preliminary interview, a patient category is assigned (see Figure 2: Patient Categories) [12]. The category is included in the database, along with a feature that lists problems in order of severity (Ex. TH is Tinnitus first, then Hyperacusis).

Category 0: Low Impact on Life, Tinnitus Present
Category 1: High Impact on Life, Tinnitus Present
Category 2: High Impact on Life, Subjective Hearing Loss Present
Category 3: High Impact on Life, Tinnitus Not Relevant, Subjective Hearing Loss Not Relevant, Hyperacusis Present
Category 4: High Impact on Life, Tinnitus Not Relevant, Hyperacusis Present, Prolonged Sound-Induced Exacerbation Present

Figure 2. Patient Categories

Counseling begins immediately and all information on patients is tied to a patient id, visit number, and visit date. During every visit, patients complete a visit questionnaire referred to as a Tinnitus Handicap Inventory. This questionnaire provides a self assessment of patient treatment progress related to emotional and other measures. Additionally, patients

in the new dataset complete a Tinnitus Function Inventory which has features related to cognitive and emotional aspects of the patient that are affected by tinnitus. Instruments (table top sound generator, in ear sound generator) are assigned and tracked. A follow-up questionnaire in the same form as the original Interview Questionnaire is administered at or near the end of treatment.

2.3 Database Features

A tinnitus patient database of ten tables and 555 patient tuples was prepared at Emory University School of Medicine. A second database of eleven tables and 758 tuples was also prepared by Dr. Jastreboff's Center; this dataset includes one additional table containing the patient scores from the new Tinnitus Functional Inventory. All identifying information related to the patient has been removed from both databases in keeping with privacy laws.

Figure 3 shows all tables and original attributes (including the new TFI table in Figure 4) and will be used as a basis for discussion of the features in the dataset.

Demographic	
PK	THC #
	Date G DOB State Zip Country T Induced H Induced Comments

Neuman_Q	
PK,I1 PK	THC # Date
	v # F-1 F-2 E-3 F-4 C-5 E-6 F-7 C-8 F-9 E-10 C-11 F-12 F-13 E-14 F-15 E-16 E-17 F-18 C-19 F-20 E-21 E-22 C-23 F-24 E-25 Sc F Sc E Sc C Sc T Comments

Questionnaires_tin	
PK,I1 PK	THC # Date
	v # Prob C Ms CC Instr Where > Fluc Sleep h Sleep desc Conc Sleep QRA Work Rest Sprt Soc Oth oth_des Aw%T An%T Tch T Sv T An T EL On perc On prbl On G/S On assoc Bad D Freq As Freq As Bad Eff snd How Ing Treatm Why prob Comments

Questionnaires_HL	
PK,I1 PK	THC # Date
	v # Hp HA HAt HAr Com Out T pr HL pr Pr Ret Recom Next v Next t Comments

REM	
THC # date	
	Freg RE Th R SPL Mix R SPL Mix R SL Tol R SPL Tol R SL Max R SPL Max R SL Freg LE Th L SPL Mix L SPL Mix L SL Tol L SPL Tol L SL Max L SPL Max L SL Category Instruments Comments

Questionnaires_DST	
PK,I1 PK	THC # Date
	v # DST Hp Phys Desc Concert shopp Mov Wrk Rest Drv Sport Church House Child Soc Oth Oth_des H Sv H An H EL Bad D Freq As Freq As Bad Eff snd How Ing Prot %T When Treatm Why prob Comments

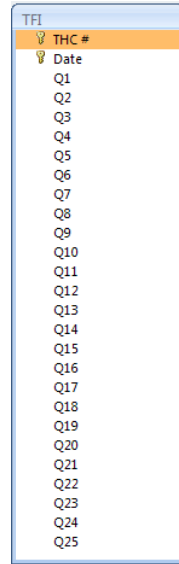
Pharmacology	
PK,I1 PK PK	THC # v # Med #
	Medication Generic Dose Duration Cat chem Action application Usual MAXim T side Comments

Instruments	
PK,I2 PK	TH# Date
I1	v # Ins Type Model ID Comments

Audiological	
PK,I1 PK	THC # Date
	v # R TD 1 R TD f 2 R TD 2 R TD f 3 R TD 3 L TD 1 L TD f 2 L TD 2 L TD f 3 L TD 3 R25 R50 R1 R2 R3 R4 R6 R8 R10 R12 L25 L50 L1 L2 L3 L4 L6 L8 L10 L12 T PR T Rm T LR Th R T RLs T PL T Lm T LL Th L T Ls WNR WNL MRR MRL MRB MLR MLL MLB MBR M BL M BB R SD L SD LR50 LR1 LR2 LR3 LR4 LR6 LR8 LR12 LRTP LL50 LL1 LL2 LL3 LL4 LL6 LL8 LL12 LLTP Comments

Miscel	
I1	ID v # Ed Deg Occup Work Comments

Figure 3: Original Database Description (from Access)



THC #
Q1
Q2
Q3
Q4
Q5
Q6
Q7
Q8
Q9
Q10
Q11
Q12
Q13
Q14
Q15
Q16
Q17
Q18
Q19
Q20
Q21
Q22
Q23
Q24
Q25

Figure 4: New Tinnitus Functional Index Table

The database in the original form is a third normal form relational database, and the metadata is enhanced to include a comment on each attribute explaining the contents.

The demographic table contains features related to gender, date of birth, state and zip. The tuples of the demographic table are uniquely identified by patient id, and one tuple per patient exists. Additionally, three text attributes are present that contain information on how and when the tinnitus and hyperacusis were induced, and a comments attribute that contains varied information that may be of interest to the research. Text fields such as these required some work before they can be used. In the original state, they were not useful as the information was hidden in the narrative. Further complications existed due to misspellings, missing values, and inconsistencies in the way information was represented. For example, it is of interest in continuing research to separate patients whose tinnitus was induced by a loud noise. A new Boolean feature was developed that shows if the tinnitus was induced in this way or not. In order to create this attribute, the T-induced, H-induced, and Comments attributes from the Demographic table needed to have the text mined while

looking for key words that are derived from the domain knowledge. Key words for this task include “loud noise”, “concert”, “military explosion”, etc. If these words are present, the Loud Noise Boolean attribute contains true. Other text mining applications show promise, and can be used to generate new rules. The occupation of the patient appears in the Comments attribute and mining this information may be relevant to new rule generation. Keywords to use in mining will need to be developed, and may be used to create an additional Boolean field related to whether the patient is in a professional type position or not. Additionally, medications that the patient has or is taking show interest as they affect the treatment process and success.

The miscellaneous table contains patient id and visit number. This table stores information on patient occupation, highest educational degree, and a comments attribute that presents interesting possibilities once again for text mining. The comments attribute contains information such as “speaks Spanish”. Future work may include text mining on this field to allow inclusion in the knowledge discovery process, with the hopes of additional rule generation.

The Neumann-Q table stores the data from the Tinnitus Handicap Inventory. This inventory is extremely important to mapping treatment progress. Information stored in the table represents patients responses to questions related to their tinnitus effect on their functioning (F), emotions (E), how catastrophic it is (C), and then a total score (T) is calculated by adding the F, E, and C scores. The total score (T score) is important as it is a measure of tinnitus severity. T score of 0 to 16 represents slight severity, 18 to 36 is mild, 38 to 56 is moderate, 58 to 76 is severe, and 78 to 100 is catastrophic [10]. The Tinnitus

Handicap Inventory is completed during each patient visit and stored with Patient ID, Visit Number and Date. These attributes can be used in a relationship to other tables.

The Pharmacology table once again uniquely identifies attributes by Patient ID, Visit Number and Visit Date. This table stores information on medications taken by the patient. All information is stored in text form and may be used in later research.

Three tables are used to store information from the preliminary and follow-up questionnaire: Questionnaires-DST, Questionnaires-HL, and Questionnaires-Tin. Questionnaires-DST provides the information from the questionnaire related to sound tolerance, questionnaires-HL relates to hearing loss, and questionnaires-Tin is related to tinnitus. These tables contain a tremendous amount of information and a patient will typically have an entry in each table at the beginning of treatment, with additional questionnaires represented almost every time they receive treatment. The information in the tables is identified by Patient ID, Visit Number and Date. The Visit Number is sometimes recorded as -1, meaning the questionnaire was completed before the first visit. One attribute that also is useful is Prob which shows problems in order of importance: T represents tinnitus, H represents hyperacusis, L represents hearing loss; and if no problem the letter is omitted. The attribute may contain an entry such as “TL” meaning the patient’s primarily problem is Tinnitus, followed by Hearing Loss.

The Instruments table contains information on the type of instrument prescribed to the patient. This table is identified by Patient ID, Visit Number, and Visit Date. Patients can receive more than one type of instrument during the course of treatment.

The Audiological table contains information from the various Audiological tests given during treatment. This table presented the most difficulty in understanding, as

knowledge of audiology is important particularly as the audiological testing relates to tinnitus discovery and treatment. The tuples in the table are identified by Patient ID, Visit Number and Visit Date.

The Tinnitus Functional Index table contains information regarding patients ratings relative to tinnitus effect on cognitive and emotional factors. The TFI is a new index (questionnaire) introduced to 75 unique patients that were included in the combined, new dataset. Most of the 161 TFI tuples also include a total score from the Tinnitus Handicap Inventory. Questions from the Tinnitus Functional Index are mapped to a hierarchical model that describes emotions invoked by music in which the main elements are stress and energy that represent two (out of four) the most general values of attributes.

New emotion related features E1, E2, E3 and E4 will be discussed in Chapter 3.

2.4 Extraction, Transformation and Loading of Original Features

Useful features from the original and combined database deemed pertinent to data mining were first extracted and transformed in preparation for analysis as separate datasets. The goal was to extract those features that described the situation of the patient based on the behavior of the attributes over time, and to transform, discretize and classify them in new ways that are useful, resulting in one table that could then be used in mining. Many algorithms exist for discretization, yet in this research the expert domain knowledge provided the basis for many of the discretization algorithms. This section will identify the resulting features along with a description of the transformation performed.

The patient id was standardized across tables. Patient id, along with visit number and date, is an important identifier for individual tuples and varied slightly in type and length in different tables. This was relatively easy to identify and correct. The visit number

and visit data were transformed to total visits (representing number of visits) and length of visit, a continuous temporal feature that determined time span in number of days between first and last visits.

Patient data related to visits includes a determination of the problem in order of importance, stored as various combinations of the letters “T” for tinnitus, “H” for Hyperacusis, and L for “Loudness Discomfort”. Only the first and last of these (First P and last P) are stored in two separate attributes in the analysis table related to first and last visit. Using the first and last gives the researchers information on problem determination at the beginning of the treatment cycle, and at the end when the patient should have moved toward category 0, indicating successful treatment.

Patient category represents the classification of the patient and is represented twice: first by original category as previously described, and second by category of treatment prescribed by the treatment specialist. To review, this feature is represented by a range of scores from 0 to 4 where 0 represents tinnitus as a minimal problem, 1 represents tinnitus as a significant problem, 2 represents tinnitus as a significant problem and hearing loss a significant subjective problem, 3 represents tinnitus as irrelevant and hyperacusis as a significant problem with hearing difficulties irrelevant, and 4 represents prolonged tinnitus with hearing difficulties irrelevant [11]. Two patient categories are stored in the final table (C and Cc), the first category assigned representing the diagnosis and the last category assigned representing the final determination of the patient problem. Assigning the patient to a category is important to treatment success, and a successfully treated patient will move toward category 0 [6]. Some analysis was performed based on this feature.

The Tinnitus Handicap Inventory score (T Score) was discretized based on the domain knowledge. Overall, the total score represents the sum of the Score for Emotions, Function, and Catastrophic areas of the questionnaire. Lower Total T Scores are better. The difference in T Score from first to last visit was calculated and discretized to represent the improvement in the patient with categories from “a” to “e”, with “a” for good to “e” for bad. This feature is stored as Category of T score. [13]

The standard deviation of audiological testing features related to loudness discomfort levels was derived and stored in various attributes in the analysis table. Loudness discomfort level is a measure of decreased sound tolerance as indicated by hyperacusis or discomfort to sound, misophonia or dislike of sound and phonophobia or fear of sound. Expert knowledge indicates that loudness discomfort levels change with treatment and patient improvement, unlike other audiological features. Normal loudness discomfort levels are 90 – 100 dB with 102 being average normal. People with decreased sound tolerance average 81.7 dB. [14] [15] For this reason the audiological data related to loudness discomfort levels is included in analysis.

Finally, information on instruments and models of equipment used by the patient is stored in text format in the analysis table. Expert knowledge indicates that the type of the instrument is the most important feature. [14]

In preparation for mining, databases were flattened with each tuple representing a single patient history record.

CHAPTER 3: NEW FEATURE CONSTRUCTION FOR THE TINNITUS DATABASE

3.1 Text Extraction and Mining

Many of the features in the original and combined database that are stored in text format contain important information which may have correlation to features indicating treatment success, such as the Total Score from the Tinnitus Handicap Inventory. Text features related to the cause of tinnitus are of particular interest, including Boolean features describing if the patient has stress, if they take medication for depression, and if their tinnitus was caused by a loud noise.

Text mining (also referred to as text classification) involves identifying the relationship between business categories and the text data (words and phrases). This allows the discovery of key terms in text data and facilitates automatic identification of text that is “interesting”. Originally, SQL Server Integration Services, Transact SQL and VBA were used to extract terms from the text columns of T-induced, H-induced, and Comments of the Demographic table. The goal was to create new Boolean features that indicate the cause of tinnitus. Initially, work involved determining if tinnitus was induced by exposure to a loud noise. The following are the text mining steps that were used on the original database; the knowledge gained from this process was used to continue mining the 758 new tuples for these new Boolean features:

1. Term extraction transformation was used which performs such tasks as Tokenizing Text, Tagging Words, Stemming Words, and Normalizing Words. By this transformation, 60 frequent terms were determined from the T-induced feature (which is a text feature that describes how tinnitus was induced).

2. After reviewing these terms, some terms were determined to be inconsequential in the domain. These terms were classified as noise words as they occurred with high frequency. These terms were then added to the exclusion terms list which is used as a reference table for the second run of the Term extraction transformation.
3. The second Term extraction transformation resulted in 10 terms which are related to the tinnitus induced reason of “noise exposure”. These terms were used to make up the dictionary table.
4. Fuzzy Lookup transformation was applied which uses fuzzy matching to return close matches from the dictionary table to extract keywords/phrases into the new Boolean feature “IsNoiseExposure”. This attribute indicates whether the induced reason for tinnitus is related to exposure to a loud noise of some type.
5. After adding this new attribute to the table, data mining algorithms (Decision Tree) were applied in order to produce relevant rules. In the original database, twenty-nine patients have the value of true for the new attribute “Noise”, and these are identified by Patient ID.

Similar work was completed to develop the new attributes for cause of tinnitus relating to stress (Stress) and medical reasons (Medical). Table 1 shows the key words used to develop the new Boolean features for Stress, Noise, and Medical. The features are sparsely represented in the database with stress appearing in 7 out of 253 patients, noise appearing in 29 out of 253 patients, and medical appearing in 22 out of 253 patients in the original dataset.

Table 1: New Boolean Features

New Boolean Features Stress, Noise, and Medical Based on Text Mining of Terms	
Stress	stress, depression, emotion, work, marriage, wedding
Noise	accident, noise, concert, loud, music, shooting, blast
Medical	surgery, infection, medicine, depression, hospital

3.2 Temporal Feature Development and Extraction

Temporal features have been widely used to describe subtle changes of continuous data over time in various research areas, such as stream tracer study [16], music sound classification [17], and business intelligence [18]. It is especially important in the light of the tinnitus treatment process. Evolution of sound loudness discomfort level parameters in time is essential for treatments; therefore it should be reflected in treatment features as well. The discovered temporal patterns may better express treatment process than static features, especially considering that the standard deviation and mean value of the sound loudness discomfort level features can be very similar for sounds representing the same type of Tinnitus treatment category, whereas changeability of sound features with tolerance levels for the same type of patients makes recovery of one type of patients dissimilar. New temporal features include:

$$C = \frac{\sum_{n=1}^{length(T)} n / length(T) \cdot V(n)}{\sum_{n=1}^{length(T)} V(n)}$$

Figure 5: Sound Level Centroid

This feature is represented as C is the gravity center of the sound level feature V, V(n) is the value of the sound level feature V in the nth visit, and T is the total number of visits.

An example would be a patient with three total visits represented by $T = 3$, and V is represented by an improving Loudness Discomfort Level measured at each of three visits with a value of 80, 90, and 100. Sound level Centroid would be calculated as $(1/3 * 80) + (2/3 * 90) + (3/3 * 100)$ divided by $(80 + 90 + 100)$ giving a result of .685185 for the Sound Level Centroid.

$$S = \sqrt{\frac{\sum_{n=1}^{\text{length}(T)} (n / \text{length}(T) - C)^2 \cdot V(n)}{\sum_{n=1}^{\text{length}(T)} V(n)}}$$

Figure 6: Sound level Spread

In this new feature, C represents the Sound Level Centroid for the patient visit sequence.

Given the same patient and visit sequence from the previous example, Sound level spread would be calculated as the square root of $80 * (.33 - .685185)^2 + 90 * (.66 - .685185)^2 + 100 * (1 - .685185)^2$ divided by the sum of 80, 90, and 100 representing the sound feature measured at each of three visits. The result value for Sound Level Spread is 0.272576.

$$R = \frac{TS(0) - TS}{D(k) - D(0)}, \text{ where } TS = \min\{TS(i) : i \in [0, N]\}$$

k is defined as the smallest $n \in [0, N]$ such that $TS = TS(n)$

Figure 7: Recovery Rate

In Recovery Rate, TS represents the total score from the Tinnitus Handicap Inventory in a patient visit. TS(0) is the first score recorded from the Inventory during the patient initial visit. TS(k) represents the minimum total score which is the best out of the vector of the scores across visits. TS(0) should be greater meaning the patient is worse based on the Inventory from the first visit. D(k) is the date that has the minimum total score, D(0) is the date that relates to TS(0).

For the same patient example, if the first total score from the THI is 86 with a visit date of 01/01/2008 and the minimum total score is 48 recorded at a visit date of 04/01/2008,

then recovery rate is equal to $(86 - 48)$ divided by the difference in days of 91 resulting in a value of .417582. A large recovery rate score can mean a greater improvement over a shorter period of time. XY scatter plots were constructed for the original database using recovery rate compared to patient category, and recovery rate compared to treatment category in order to examine interesting patterns in the data.

3.3 Feature Development for Categorical Data

During a period of medical treatment, a doctor may change the treatment from one category to another based on the recovery of the patient. Also, the symptoms of a patient may vary as a result of the treatment; therefore, the category of patient may change over time. Other typical categorical features in our database include instruments in each treatment as well as visit dates. Statistical and econometric approaches to describe categorical data have been well discussed by Daniel Powers and Yu Xie [19]. Most frequent pattern MFP counts the pattern, which occurred most frequently for a particular patient. First and last pattern FP/LP represents the initial and final state of a categorical attribute respectively.

In the tinnitus database, statistical features such as the most frequent pattern, the first pattern and the last pattern were used to describe the changes of categorical data over time. Specifically, the problem representing the Patient problem category was represented as the most frequent pattern. The problem is a category representing a combination of T for Tinnitus, H for Hyperacusis, and L for Loudness Discomfort with the most important problem being listed first and other problems listed in decreasing order of importance.

3.4 New Features based on the Tinnitus Functional Index and Emotions

The extended dataset received late in the research contains a new table representing a Tinnitus Function Index questionnaire. Research on this questionnaire started as early as 2004 by Mary B. Meikle and a large group of additional clinical investigators. The TFI has potential as becoming the primary outcome measure for treatment of tinnitus [20]. The TFI represents 25 questions, with 24 being rated on a numeric rating scale from 0 to 10 in an 11 point scale. An example of this type of scale is found in Figure 8:

In the question below, please circle the number that best describes you:

Over the past week, how ANXIOUS has your tinnitus made you feel?

0 1 2 3 4 5 6 7 8 9 10

Not at all

Extreme

(reference needed Tinnitus Outcomes Assessment Meikle et al 231).

Figure 8: Sample of TFI Question

The 25 questions (only question 1 is scored on a percentage basis) are represented in Table 2 below. This research also mapped the questions to a Category of Question based on the description of the question. Of particular interest are the categories for E1, E2, E3, and E4 representing emotional categories related to the Emotional-Valence Plane developed by Thayer [7]. The questions are mapped to E1 Energetic Positive, E2 Energetic Negative, E3 Calm Negative, and E4 Calm Positive in the Thayer model as follows in Table 2.

Table 2: Tinnitus Functional Index (scale of 0 to 10)

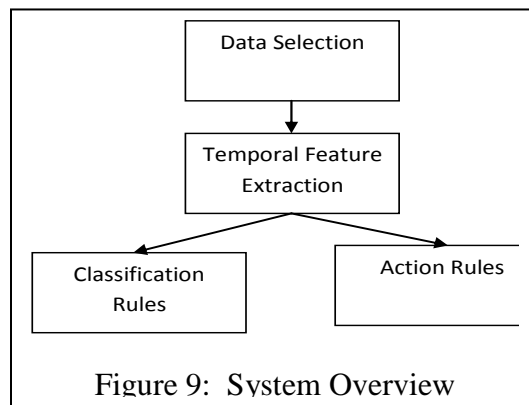
		Category of Question	E-V Scale
Q1	% of time aware	Awareness	
Q2	loud	HEARING	
Q3	in control	E11	E1
Q4	% of time annoyed	Annoyance	
Q5	cope	E11	E1
Q6	ignore	E21	E2
Q7	concentrate	THINKING CONCENTRATION	
Q8	think clearly	THINKING CONCENTRATION	
Q9	focus attention	THINKING CONCENTRATION	
Q10	fall/stay asleep	E33	E3
Q11	as much sleep	E33	E3
Q12	sleeping deeply	E33	E3
Q13	hear clearly	HEARING	
Q14	understand people	HEARING	
Q15	follow conversation	HEARING	
Q16	quite, resting activities	E41	E4
Q17	relax	E43	E4
Q18	peace and quiet	E42	E4
Q19	social activities	SOCIAL	
Q20	enjoyment of life	E11	E1
Q21	relationships	SOCIAL	
Q22	work on other tasks	SOCIAL	
Q23	anxious, worried	E23	E2
Q24	bothered upset	E22	E2
Q25	depressed	E31	E3

Sum of values represents E1 Energetic Positive, E2 Energetic Negative, E3 Calm Negative, E4 Calm Positive

The score in the related category is summed, representing the new attribute E1, E2, E3 and E4. In total, 136 patient visit tuples are represented with the TFI questionnaire; most of these also completed the THI with only the total score recorded in the extended dataset for these tuples.

CHAPTER 4: CLUSTERING TECHNIQUES FOR FEATURE EXTRACTION

Tinnitus patients exhibit patterns in visit frequency and this can be used to group patients for more effective decision making related to treatment. A flexible temporal feature retrieval system is developed as a part of this research. The system is based on grouping the patients of similar visiting frequencies with connection to classification-rules discovery engine an action-rules discovery engine, which consists of four modules: a data grouping device, a temporal feature extraction engine, classification rules generation device, and an action rules generation device [21]. See Figure 9 for System Overview.



The data grouping device is to filter out less relevant records in terms of visiting duration patterns measured from the initial visits. The temporal feature extraction engine is to project temporal information into patient-based records for classic classifiers to learn effects of treatment as well as tinnitus upon patients. WEKA (J48, Random Forest, and Multilayer Perceptron (Weka's implementation of Neural Networks)) are used to build and evaluate the classifiers to be used by Decision Support System for Tinnitus. To extract action rules, a new Frequent-Sets based action rules generator has been built and

implemented, (conceptually similar to the system proposed in [22]) and also used was Jan Rauch's Lisp_Miner [23] [24] [25]. This data analysis has been performed on the original database extended by new temporal features developed for the clustered data and related to the visit sequence and Total Score.

4.1 Data selection, Grouping, and Temporal Feature Extraction

Tuples are grouped by similar visiting patterns, where the visiting history of each patient is discretized into durations, anchored from its initial visit date in terms of weeks and serving as a seed for grouping. This process was applied to the original database in preparation for grouping data into frequent visit sets for classification rule discovery. For example, a patient p who visited a doctor on July 8th, 2009, August 14th, 2009, and October 7st, 2009 is recorded as Table 1.

Table 3: An example of calculating visit duration

Visit ID	Duration (weeks)
1	6
2	14

The corresponding vector representation will have the form $v_p = [6, 14]$. It means that patient p visited the doctor five full weeks after his first visit and his last visit happened 13 weeks after his first visit (or 7 weeks after his second visit). In other words, patient p visited the doctor in the 6th week and 14th week in the relation to his first visit. Assume now that we have two patients denoted by p, q . Patient p visits are represented by a vector $v_p = [v_1, v_2, \dots, v_n]$ whereas vector $v_q = [w_1, w_2, \dots, w_m]$ represents visits of patient q . If $n \leq m$, then the distance $\rho(p, q)$ between p, q and the distance $\rho(q, p)$ between q, p is defined as

$$\rho(q, p) = \rho(p, q) = \frac{\sum_{i=1}^n |v_i - w_{J(i)}|}{n}, \quad (1)$$

where $[w_{J(1)}, w_{J(2)}, \dots, w_{J(n)}]$ is a subsequence of $[w_1, w_2, \dots, w_m]$ such that

$$\sum_{i=1}^n |v_i - w_{J(i)}|$$

is minimal for all n -element subsequences of $[w_1, w_2, \dots, w_m]$. By $|v_i - w_{J(i)}|$

we mean the absolute value of $[v_i - w_{J(i)}]$.

For example, if patient p , having four visits, is compared to patient p' , who had five visits; each of the three visits (second, third, fourth) of patient p shall be matched with a closest visit of p' and their difference shall be averaged.

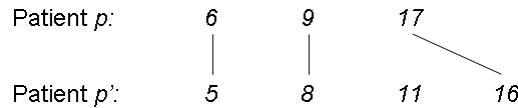


Figure 10: Matching for Closest Visit Pattern

In the example shown in Figure 2, the distance $\rho(p, p') = 1$.

It can be easily checked that $\rho(q, p)$ is reflexive and symmetric but not transitive which means it is a tolerance relation.

A threshold is applied to filter out patient records with large distance values to form a tolerance class, where all group members have similar visiting patterns; therefore visit-related temporal features can be computed for all group members.

For instance, let us assume that we have 8 patients $p_1, p_2, p_3, \dots, p_8$ with doctor's visits assigned to them which are represented by vectors:

$$v_{p1} = [3, 8, 12, 20], v_{p2} = [4, 7], v_{p3} = [5, 12, 21, 30], v_{p4} = [7, 21, 29],$$

$$v_{p5} = [12, 22], v_{p6} = [13, 19, 29], v_{p7} = [2, 13, 19, 31, 38], v_{p8} = [7, 12, 20].$$

The threshold value $\rho=1$ is set up as a minimal distance between vectors representing patients.

The following tolerance classes containing more than one element are constructed:

$$TC_{\rho=1}(v_{p2}) = \{v_{p1}, v_{p2}\}, TC_{\rho=1}(v_{p4}) = \{v_{p4}, v_{p3}\}, TC_{\rho=1}(v_{p5}) = \{v_{p5}, v_{p1}, v_{p3}, v_{p8}\}, TC_{\rho=1}(v_{p6}) = \{v_{p6}, v_{p7}\}, TC_{\rho=1}(v_{p6}) = \{v_{p6}, v_{p1}\}.$$

We say that $TC_{\rho=1}(v_{p2})$ is generated by $p2$ and similarly $TC_{\rho=1}(v_{p4})$ is generated by $p4$.

The ultimate goal of constructing tolerance classes is to identify the right groups of patients for which useful temporal features can be built and used to extend the database. By increasing the threshold value, we get larger classes for the process of knowledge extraction, but the information included in temporal features will be less accurate. On the other hand, if the threshold value is too small, the size of tolerance classes might be also too small in order to get any useful information through the knowledge extraction process.

4.2 Temporal feature extraction with Clustered Data.

The dataset associated with a tolerance class which is generated by patient p contains records describing patients who visited their doctor at least during similar weeks as the patient p . Data referring only to these visits are stored in tuples representing all patients in this tolerance class. In other words, if patient p generates a tolerance class $TC_{\rho=1}(v_{p2})$ where $v_{p2} = [4, 7]$ and another patient $p1$ has a vector representation $v_{p1} = [3, 8, 12, 20]$ of his doctor's visits, then $p1$ has a vector representation $[3, 8]$ relative to $TC_{\rho=1}(v_{p2})$. This way all

patients associated with the same tolerance class have the same number of doctor's visits and all these visits happened approximately with the same visit distance.

The construction of a collection of databases D_p was performed, where p is a patient and D_p corresponds to $TC\rho(v_p)$, for the purpose of classifiers construction based on the tinnitus database O that was mentioned in the previous section. The term "attribute" is used to refer to a column in the table from the database O and the term "feature" to refer to a column in the database D_p . Also, due to the intuition of the process in each visit, recovery of any patient with only one visit cannot be evaluated. Therefore, such records have been removed during the experiments. During a period of medical treatment for tinnitus, a doctor may change the treatment from one category to another based on the specifics of recovery of the patients and the symptoms of a patient may vary as a result of the treatment. Additionally, the category of patient may change over time (e.g., hyperacusis can be totally eliminated and consequently the patient may move from treatment category 3 to 1). Other typical categorical features which may change over time in the database O include sound-instrument types as well as visiting frequencies. Statistical and econometric approaches to describe categorical data have been well discussed in [19].

In terms of continuity, there are two types of data: one is numerical, such as scores for emotions, functions, and catastrophes related to the tinnitus problems; the other is categorical, such as instruments used in the therapy and patient categories. In terms of stability, there are two other types of data: one is stable; the other is flexible [26]. In this research, stable is defined relative to others: an attribute should have the same value along time throughout the most of the records (some threshold is given).

Under all the above assumptions, the transformation from visit-based format O to patient-based format Dp for each tolerance class TCp(vp) is quite straightforward.

Let us assume that TC(vp) is a tolerance class generated by patient p where vp = [v₁, v₂, ..., v_n] and [w_{J(1)}, w_{J(2)}, ..., w_{J(n)}] is a vector representation of patient q relative to TC(vp). We also assume that J(0)=1.

Now, assume that A is a numerical attribute which time-dependent values for patient q are given as a vector [a_{J(1)}, a_{J(2)}, ..., a_{J(n)}]. For each patient q ∈ TCp(vp), if n is an even number, we compute the temporal feature value A1(q) to describe the derivative of an attribute A against a number of rounded weeks between his doctor's visit J(0) and J(n)/2. We also compute A2(q) to describe the derivative of an attribute A against a number of rounded weeks between his doctor's visit J(n)/2 and J(n). Finally, we compute A3(q) to describe the derivative of an attribute A against a number of rounded weeks between his doctor's visit J(0) and J(n).

$$A_1(q) = \frac{[a_{J(n)/2} - a_{J(0)}]}{|w_{J(n)/2} - w_{J(0)}|} \quad (2)$$

$$T_1(q) = [a_{J(n)/2} - a_{J(0)}] \quad (3)$$

$$A_2(q) = \frac{[a_{J(n)} - a_{J(n)/2}]}{|w_{J(n)} - w_{J(n)/2}|} \quad (1)$$

$$T_2(q) = [a_{J(n)} - a_{J(n)/2}] \quad (2)$$

$$A_3(q) = \frac{[a_{J(n)} - a_{J(0)}]}{|w_{J(n)} - w_{J(0)}|} \quad (3)$$

$$T_3(q) = [a_{J(n)} - a_{J(0)}] \quad (4)$$

When n is an odd number, we developed the temporal feature A4(q) to describe the derivative of an attribute against time of rounded week of its first visiting duration.

Temporal feature A5(q) is similar to A3(q).

$$A_4(q) = \frac{[a_{J(1)} - a_{J(0)}]}{|w_{J(1)} - w_{J(0)}|} \quad (5)$$

$$T_4(q) = [a_{J(1)} - a_{J(0)}] \quad (6)$$

$$A_5(q) = \frac{[a_{J(n)} - a_{J(1)}]}{|w_{J(n)} - w_{J(1)}|} \quad (7)$$

$$T_5(q) = [a_{J(n)} - a_{J(0)}] \quad (8)$$

New features, A6(q) and T6(q) are defined similarly to A1(q) and T1(q). Finally, A7(q) and T7(q) are defined similarly to A2(q) and T2(q).

4.3 New Temporal Features for Clustered Visits: Coefficients and Angles

The clustering algorithm provided a collection of databases for three and four visit sets based on distance determined by a patient seed tuple. From these three and four visit sets, information on the distance between visits and a database feature (Total Score from the Tinnitus Handicap Inventory) is used to develop a unique new set of temporal based features developed by the coefficients of the polynomial equation that best represents the visits and by the angles that are formed from the plot on the line for visit length (x axis) and score (y axis). The analysis of the new features representing angles developed for four visit clustered sets is as follows:

$$\theta_{v_i \rightarrow v_j} = \tan^{-1} \left(\frac{ds_{v_i \rightarrow v_j}}{dt_{v_i \rightarrow v_j}} \right) (180\pi)$$

$\theta_{v_i \rightarrow v_j}$ is calculated by the equation above in terms of $ds_{v_i \rightarrow v_j}$ and $dt_{v_i \rightarrow v_j}$

$ds_{v_i \rightarrow v_j}$ is the difference of the total score between visit v_i and visit v_j

$dt_{v_i \rightarrow v_j}$ is the difference of the time in weeks between visit v_i and visit v_j

The following features are added:

$$\theta_{v_1 \rightarrow v_2} \theta_{v_1 \rightarrow v_3} \theta_{v_1 \rightarrow v_4} \theta_{v_2 \rightarrow v_3} \theta_{v_2 \rightarrow v_4} \theta_{v_3 \rightarrow v_4} ds_{v_1 \rightarrow v_2} ds_{v_1 \rightarrow v_3} ds_{v_1 \rightarrow v_4} ds_{v_2 \rightarrow v_3} ds_{v_2 \rightarrow v_4} ds_{v_3 \rightarrow v_4}$$

New features are also added representing the tangents of the angles formed by the visit sequences. Angles formed for four visit sets include those between visit 1 and 2, visit 1 and 3, and visit 1 and 4.

See Figure 11: Angle Formulation with calculations between visit 1&2, 1&3, and 1&4 below.

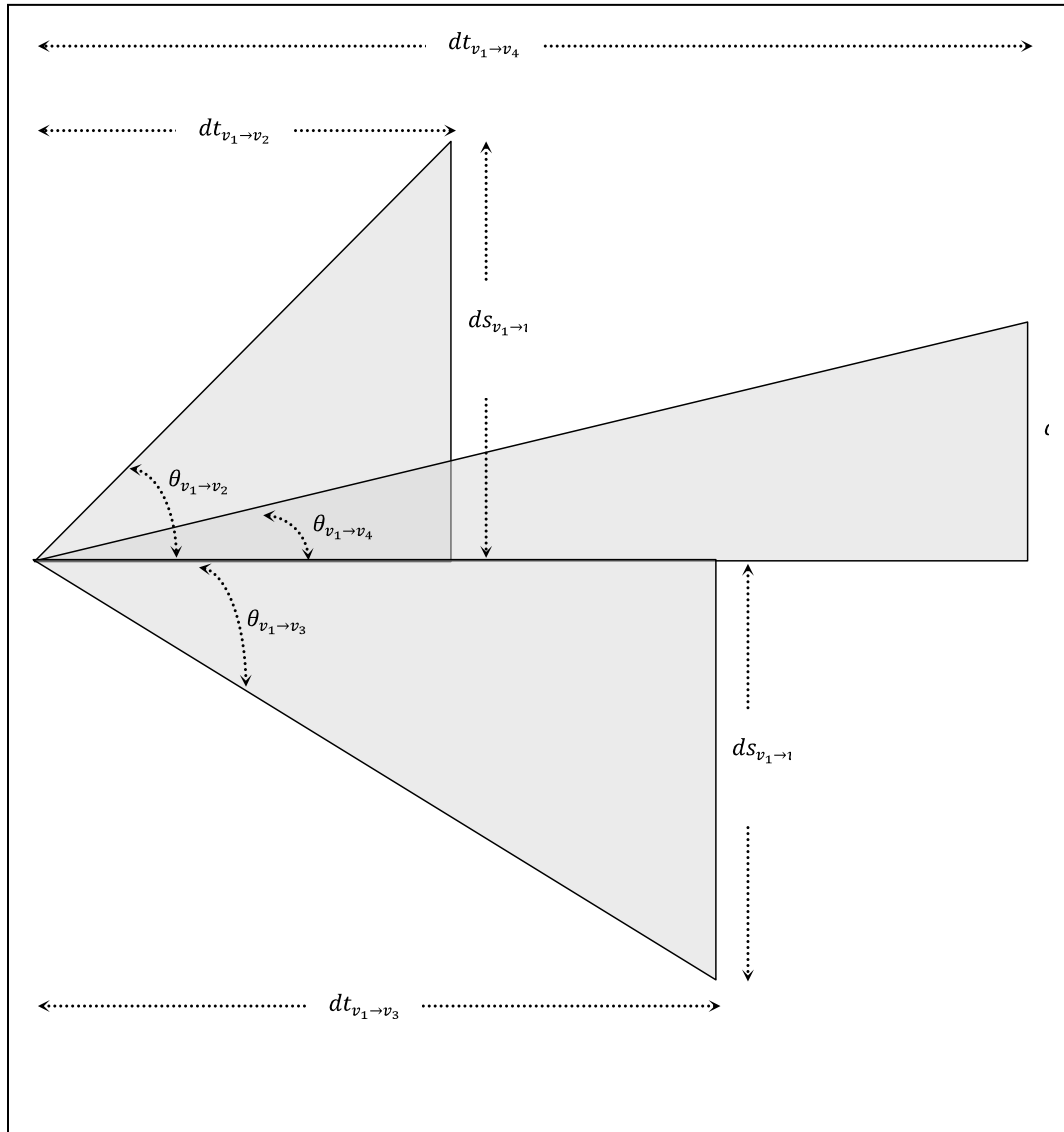


Figure 11: Angle Formulation

4.4. Quadratic Equation Based New Features

Given four points created by $(dt_{v_1 \rightarrow v_1}, s1)(dt_{v_1 \rightarrow v_2}, s2)(dt_{v_1 \rightarrow v_3}, s3)(dt_{v_1 \rightarrow v_4}, s4)$, a quadratic equation can be derived and the coefficients for the curve are added as new features.

Given $dt_{v_1 \rightarrow v_1} = 0$ and $(dt_{v_i \rightarrow v_j})^0 = 1$, we have the following set of simultaneous equations:

$$c_0 = s_1$$

$$c_0 + c_1(dt_{v_1 \rightarrow v_2})^1 + c_2(dt_{v_1 \rightarrow v_2})^2 + c_3(dt_{v_1 \rightarrow v_2})^3 = s_2$$

$$c_0 + c_1(dt_{v_1 \rightarrow v_3})^1 + c_2(dt_{v_1 \rightarrow v_3})^2 + c_3(dt_{v_1 \rightarrow v_3})^3 = s_3$$

$$c_0 + c_1(dt_{v_1 \rightarrow v_4})^1 + c_2(dt_{v_1 \rightarrow v_4})^2 + c_3(dt_{v_1 \rightarrow v_4})^3 = s_4$$

New features result from the solution for the coefficients c_0 , c_1 , c_2 , and c_3 .

Please reference Figure 12 for a diagram showing the points included in the quadratic equation.

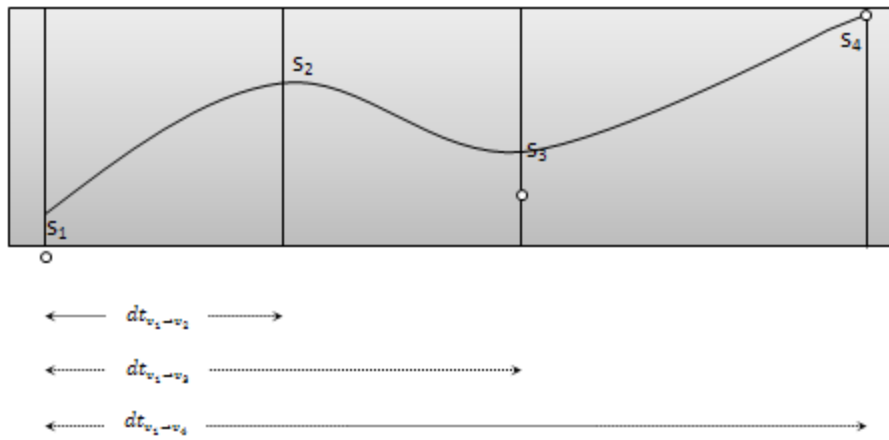


Figure 12: Points that Make Up The Quadratic Equation

CHAPTER 5: MINING UNCLUSTERED DATA

Decision tree study was initially performed using WEKA and J48, WEKA's implementation of the C4.5 decision tree learner [11], a system that incorporates the ID3 algorithm for decision tree induction. J4.8 includes improved methods for handling numeric attributes and missing values, and generates decision rules from the trees. [15]. To build the classifiers from unclustered data we continue to use WEKA and Random Forest along with Multilayer Perceptron with the discretized Total Score from the Tinnitus Handicap Inventory. The new features and discretized total score were analyzed in order to improve the confidence of the classifiers built from the tinnitus database from the original data without new features.

Random forest is an ensemble classifier that consists of many decision trees and outputs the class value that occurs most frequently as the class's output by individual trees.

A multilayer perceptron (MLP) is a feedforward artificial neural network model that maps sets of input data onto a set of appropriate output. An MLP consists of multiple layers of nodes in a directed graph, with each layer fully connected to the next one. Except for the input nodes, each node is a neuron (or processing element) with a nonlinear activation function.

5.1 Original Experiment and Results

In this initial research, two different experiments were performed: Experiment#1 explored Tinnitus treatment records of 253 patients and applied 126 attributes to investigate the association between treatment factors and recovery; Experiment#2 explored 229 records

and applied 16 attributes to investigate the nature of tinnitus with respect to hearing measurements. All classifiers were 10-fold cross validation with a split of 90% training and 10% testing. WEKA (J48) was used in all cases.

Preliminary research results showed several interesting rules resulting from decision tree analysis:

5.1.1. Experiment#1:

- (Category of treatment = C1) \wedge (R50 >12.5) \wedge (R3 \leq 15) \implies improvement is neutral
The support of the rules is 10, the accuracy is 90.9%. It means that if treatment category chosen by patient is C1 then when R50 parameter is above 12.5 and average of R3 is less or equals to 15 then the recovery is neutral.
- (Category of treatment = C2) \implies good
The support of the rules is 44, the accuracy is 74.6%. It means that if category of treatment chosen by patient is C2 then Improvement is good.
- (Category of treatment = C3) \wedge (Model = BTE) \implies good
The support of the rules is 17, the accuracy is 100.0%.

3.1.2 Experiment#2:

- 40>Lr50>19 \implies Somehow has tinnitus all of the time
The support of the rules is 27, the accuracy is 100.0%. It means that if Lr50 is in range of 19 to 40, somehow the patient has tinnitus all the time, where the tinnitus may not be a major problem.

Scatter plot analysis shows when recovery rate is compared to patient and treatment category in XY scatter plot analysis, both patient and treatment category 4 shows a smaller rate of recovery value possibly indicating slower or reduced treatment success.

5.2 Structure of the Decision Attribute

In order to improve classification after the original experiments, algorithms were applied to the Total Score from the Tinnitus Handicap Inventory and used to develop eight new decision attributes TSa through TSh based on the discretization of the difference in Total Score from the first visit (high total score is typical) to the last visit (a lower score

represents improvement). A variety of discretization methods were applied including averaging and expert knowledge. The greater the difference in Total Score, the greater the improvement with a in each discretized decision attribute representing the best improvement. The decision feature was added to the flattened dataset and used to learn the value of the new features for classification with a goal of improved classification and learning from the new features. Total Score categories are represented in Table 4.

Table 4: Categories for Total Score Discretization.

Total Score Difference Discretization	Description (score a represents the highest T Score in all cases)
TSa	$a = \{s: s > 0\}$, $b = \{0\}$, $c = \{s: s < 0\}$
TSb	$a = \{s: s > 30\}$, $b = \{s: 10 < s \leq 30\}$, $c = \{s: -10 < s \leq 10\}$, $d = \{s: -40 < s \leq -10\}$, e – remaining scores
TS _c	$a = \{s: s > 28\}$, $b = \{s: 0 < s \leq 28\}$, $c = \{s: -1 < s \leq 0\}$, $d = \{s: -15 < s \leq -1\}$, e – remaining scores
TS _d	$a = \{s: s > 40\}$, $b = \{s: 10 < s \leq 40\}$, $c = \{s: -10 < s \leq 10\}$, $d = \{s: -40 < s \leq -10\}$, e – remaining scores
TS _e	$a = \{s: s > 50\}$, $b = \{s: 0 < s \leq 50\}$, $c = \{s: -50 < s \leq 0\}$, d – remaining scores
TS _f	$a = \{s: s > 80\}$, $b = \{s: 60 < s \leq 80\}$, $c = \{s: 40 < s \leq 60\}$, $d = \{s: 20 < s \leq 40\}$, $e = \{s: 0 < s \leq 20\}$, $f = \{s: -20 < s \leq 0\}$, $g = \{s: -40 < s \leq -20\}$, $h = \{s: -60 < s \leq -40\}$, i – remaining scores
TS _g	$a = \{s: s > 28\}$, $b = \{s: 0 < s \leq 28\}$, $c = \{s: -12 < s \leq 0\}$, d – remaining scores
TS _h	$a = \{s: s > 10\}$, $b = \{s: -10 \leq s \leq 10\}$, c – remaining scores

5.2.1 Extended Experiment and Results

Continuing the research, four different experiments were performed using the new decision attributes. All four experiments explored the original tinnitus treatment records of 253 patients and applied variations of 126 attributes to investigate the association between treatment factors and recovery using discretized Total Score. All classifiers were 10-fold cross validation with a split of 90% training and 10% testing. WEKA (J48) was used for all classifications.

Research results showed improved classification with several of the new features based on results from decision tree analysis (J48, Random Forest, Multilayer Perceptron) with the eight decision attributes TSa through TSh (discretized from the Total Score) and four variations of the original database representing Experiments 1 through 7 including: 1) original data with Standard Deviations and Averages from Audiological features; 2) original data with Standard Deviations, Averages, Sound level centroid and sound level spread (Sound) only; 3) original data with Standard Deviations, Averages, and Text; and 4) Original Data Standard Deviations, Averages, Text and Sound; 5) Original Data with Text; 6) Original Data with Sound; and 7) Original Data with Sound, Text, and Recovery Rate. Precision, Recall, and F-Measure were noted resulting in Tables 1 through 7 of results for each Experiment. WEKA calculates precision as the number of documents retrieved that are relevant divided by the total number that are retrieved; recall is the number of documents retrieved that are relevant divided by the total number of documents that are relevant. For example, if one system locates 100 documents and 50 are relevant as compared to another system that locates 400 documents and 60 are relevant, it is obvious that the cost of documents returned that are not relevant (false positives) and the cost of documents that are

not returned that are relevant (false negatives) is of importance. [28] Recall takes the false positives into account. F-Measure is calculated as $2 * \text{Recall} * \text{Precision} / (\text{Recall} + \text{Precision})$ and represents the harmonic mean of precision and recall. [28]

The results show that the new sound features (sound level centroid, sound level spread, and recovery rate) improve the classification result for J48, Random Forest, and Multilayer Perceptron. TSa, TSe and TSh show the best results for classification based on the discretized total score for most datasets. The WEKA results representing the best classification appear in Table 5 below.

Table 5: WEKA Results, Classifier Tree for J48		
Original Data with Sound Level Centroid, Sound Level Spread, Recovery Rate		
Decision Feature: TSa		
Precision	Recall	F-Measure
.751	.806	.776
Tree:		
Recovery Rate ≤ -0.4 : c (40.48/19.04)		
Recovery Rate > -0.4 : a (212.52/26.4)		

Figure 13 showing the WEKA results for all decision variables for the Original Data with Sound Level Centroid, Sound Level Spread, and Recovery Rate appears below:

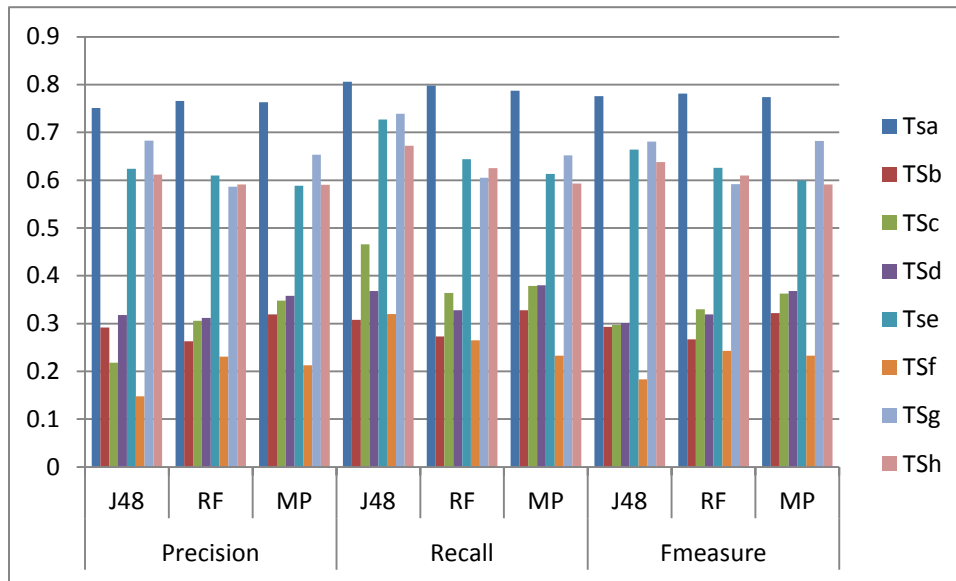


Figure 13: Top Classification Results: J48 with Decision Variable TSa and Sound Level Centroid, Sound Level Spread, and Recovery Rate

Table 6: Original Data with Standard Deviations and Averages

Original Data with Standard Deviation and Averages Only									
	Precision			Recall			F-measure		
	J48	RF	MP	J48	RF	MP	J48	RF	MP
Tsa	0.625	0.653	0.697	0.791	0.763	0.719	0.698	0.696	0.708
TSb	0.266	0.293	0.343	0.277	0.304	0.344	0.271	0.297	0.343
TSc	0.349	0.373	0.364	0.387	0.447	0.391	0.366	0.405	0.377
TSd	0.308	0.326	0.335	0.324	0.348	0.34	0.314	0.336	0.337
Tse	0.451	0.517	0.533	0.672	0.636	0.569	0.54	0.551	0.548
TSf	0.212	0.266	0.251	0.249	0.3	0.261	0.224	0.278	0.256
TSg	0.37	0.369	0.379	0.403	0.431	0.383	0.383	0.393	0.381
TSh	0.471	0.491	0.531	0.593	0.569	0.542	0.457	0.513	0.536

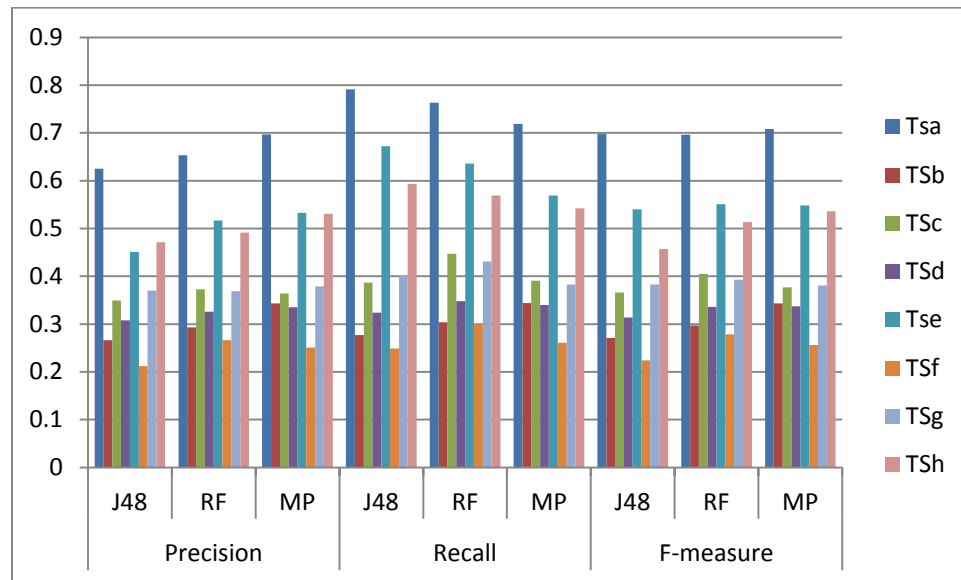


Figure 14: Graph of Table 6

Table 7: Original Data with Standard Deviations, Averages and Sound

Original Data with Standard Deviation and Averages and Sound									
	Precision			Recall			F-measure		
	J48	RF	MP	J48	RF	MP	J48	RF	MP
Tsa	0.733	0.665	0.763	0.794	0.779	0.787	0.758	0.699	0.774
TSb	0.389	0.289	0.316	0.427	0.308	0.32	0.405	0.297	0.317
TSc	0.456	0.351	0.365	0.49	0.411	0.387	0.472	0.377	0.375
TSd	0.418	0.33	0.351	0.486	0.356	0.364	0.427	0.341	0.356
Tse	0.624	0.513	0.601	0.727	0.656	0.617	0.664	0.556	0.608
TSf	0.324	0.297	0.224	0.375	0.332	0.229	0.334	0.31	0.225
TSg	0.463	0.448	0.387	0.502	0.49	0.395	0.481	0.451	0.391
TSh	0.608	0.508	0.547	0.664	0.585	0.545	0.631	0.532	0.546

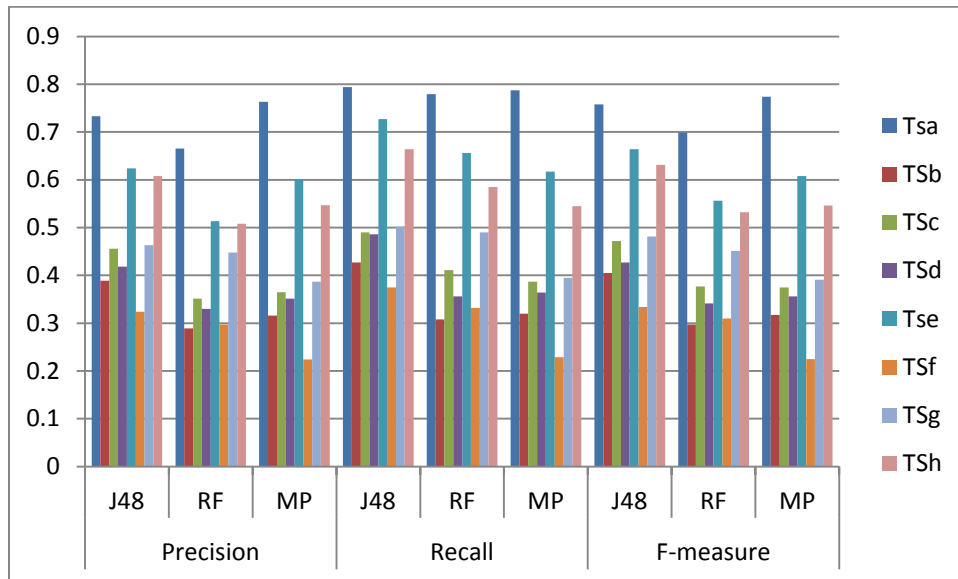


Figure 15: Graph of Table 7

Table 8: Original Data with Standard Deviations, Averages and Text

Original Data with Standard Deviation and Averages and Text									
	Precision			Recall			F-measure		
	<u>J48</u>	<u>RF</u>	<u>MP</u>	<u>J48</u>	<u>RF</u>	<u>MP</u>	<u>J48</u>	<u>RF</u>	<u>MP</u>
Tsa	0.625	0.645	0.691	0.791	0.763	0.708	0.698	0.69	0.699
TSb	0.276	0.304	0.295	0.289	0.312	0.3	0.282	0.306	0.297
TSd	0.353	0.367	0.359	0.391	0.415	0.364	0.369	0.387	0.359
TSd	0.291	0.315	0.333	0.304	0.344	0.34	0.296	0.327	0.336
Tse	0.451	0.473	0.508	0.672	0.617	0.518	0.54	0.529	0.513
TSf	0.213	0.194	0.248	0.241	0.221	0.261	0.224	0.204	0.254
TSg	0.383	0.338	0.424	0.427	0.387	0.419	0.4	0.361	0.42
TSh	0.471	0.504	0.537	0.589	0.573	0.557	0.461	0.519	0.547

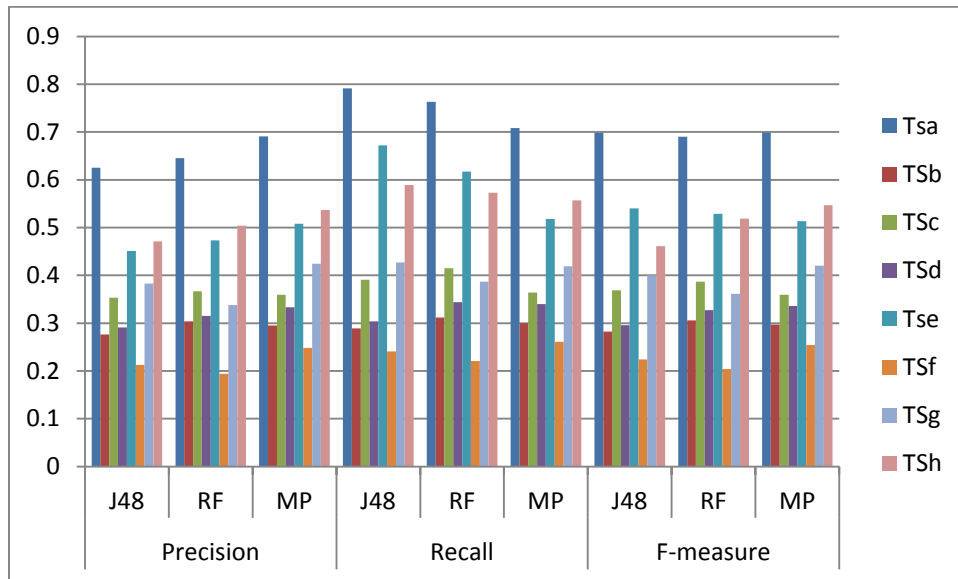


Figure 16: Graph of Table 8

Table 9: Original Data with Standard Deviations, Averages, Sound and Text

Original Data with Standard Deviation and Averages and Sound and Text									
	Precision			Recall			F-measure		
	<u>J48</u>	<u>RF</u>	<u>MP</u>	<u>J48</u>	<u>RF</u>	<u>MP</u>	<u>J48</u>	<u>RF</u>	<u>MP</u>
Tsa	0.733	0.704	0.766	0.794	0.787	0.783	0.758	0.72	0.774
TSb	0.277	0.314	0.384	0.296	0.324	0.379	0.282	0.318	0.381
TSc	0.349	0.382	0.365	0.387	0.443	0.387	0.366	0.403	0.532
TSd	0.298	0.325	0.351	0.316	0.352	0.364	0.304	0.336	0.356
Tse	0.624	0.527	0.585	0.727	0.644	0.601	0.664	0.551	0.593
TSf	0.215	0.168	0.224	0.253	0.194	0.229	0.227	0.177	0.225
TSg	0.37	0.376	0.387	0.403	0.439	0.395	0.383	0.401	0.391
TSh	0.608	0.534	0.545	0.668	0.601	0.565	0.633	0.556	0.554

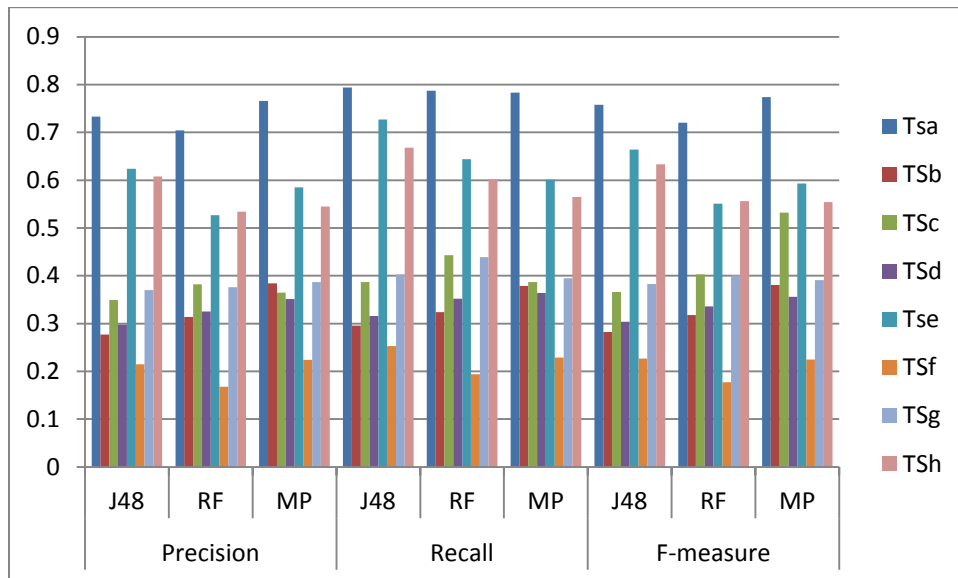


Figure 17: Graph of Table 9

Table 10: Original Data with Text

Original Data with Text									
	Precision			Recall			F-measure		
	<u>J48</u>	<u>RF</u>	<u>MP</u>	<u>J48</u>	<u>RF</u>	<u>MP</u>	<u>J48</u>	<u>RF</u>	<u>MP</u>
Tsa	0.625	0.625	0.657	0.791	0.791	0.692	0.698	0.498	0.674
TSb	0.277	0.314	0.384	0.296	0.324	0.379	0.282	0.318	0.381
TSd	0.218	0.304	0.313	0.466	0.356	0.328	0.297	0.326	0.32
TSd	0.303	0.316	0.349	0.36	0.332	0.375	0.291	0.323	0.36
Tse	0.451	0.54	0.472	0.672	0.668	0.498	0.54	0.562	0.483
TSf	0.148	0.202	0.217	0.32	0.217	0.237	0.183	0.208	0.225
TSg	0.218	0.297	0.35	0.466	0.336	0.352	0.297	0.312	0.35
TSh	0.473	0.514	0.586	0.597	0.605	0.585	0.465	0.499	0.585

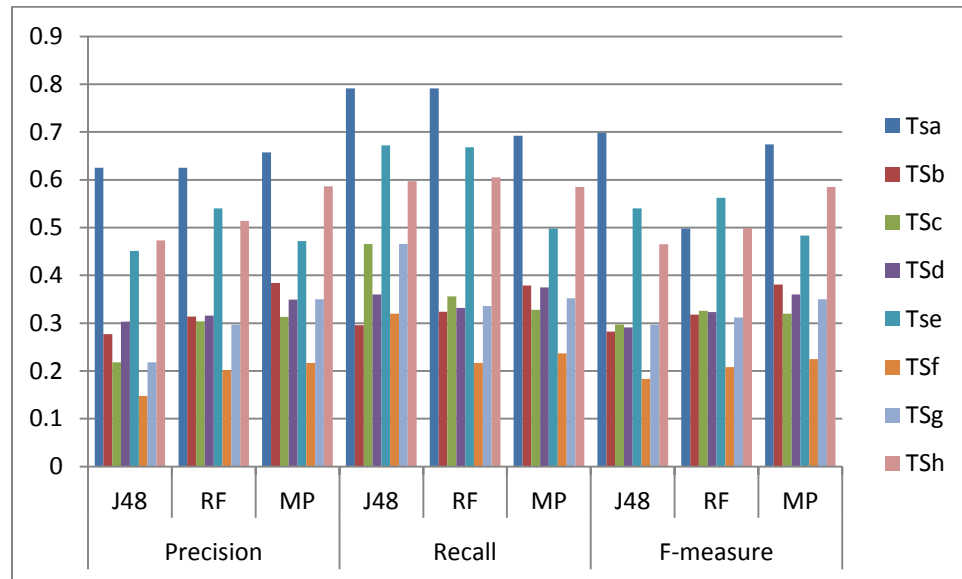


Figure 18: Graph of Table 10

Table 11: Original Data with Sound and Recovery Rate

Original Data with Sound and Recovery Rate									
	Precision			Recall			Fmeasure		
	<u>J48</u>	<u>RF</u>	<u>MP</u>	<u>J48</u>	<u>RF</u>	<u>MP</u>	<u>J48</u>	<u>RF</u>	<u>MP</u>
Tsa	0.751	0.766	0.763	0.806	0.798	0.787	0.776	0.781	0.774
TSb	0.292	0.263	0.319	0.308	0.273	0.328	0.293	0.267	0.322
TSd	0.218	0.306	0.348	0.466	0.364	0.379	0.297	0.33	0.363
TSd	0.318	0.312	0.358	0.368	0.328	0.38	0.301	0.319	0.368
Tse	0.624	0.61	0.588	0.727	0.644	0.613	0.664	0.626	0.599
TSf	0.148	0.231	0.213	0.32	0.265	0.233	0.183	0.243	0.233
TSg	0.683	0.586	0.653	0.739	0.605	0.652	0.681	0.592	0.682
TSh	0.612	0.591	0.59	0.672	0.625	0.593	0.638	0.61	0.591

(See Figure 13 above)

Table 12: Original Data with Sound, Text and Recovery Rate

Original Data Recovery Rate Sound and Text									
	Precision			Recall			F-measure		
	<u>J48</u>	<u>RF</u>	<u>MP</u>	<u>J48</u>	<u>RF</u>	<u>MP</u>	<u>J48</u>	<u>RF</u>	<u>MP</u>
Tsa	0.63	0.581	0.764	0.688	0.632	0.791	0.656	0.605	0.777
TSb	0.415	0.38	0.368	0.455	0.387	0.372	0.431	0.383	0.37
TSd	0.493	0.458	0.491	0.421	0.481	0.489	0.405	0.468	0.489
TSd	0.364	0.342	0.374	0.494	0.352	0.387	0.409	0.345	0.38
Tse	0.624	0.605	0.612	0.727	0.664	0.644	0.664	0.597	0.626
TSf	0.315	0.222	0.299	0.372	0.241	0.324	0.333	0.23	0.31
TSg	0.522	0.476	0.446	0.542	0.506	0.458	0.522	0.488	0.451
TSh	0.612	0.562	0.601	0.672	0.613	0.609	0.638	0.584	0.604

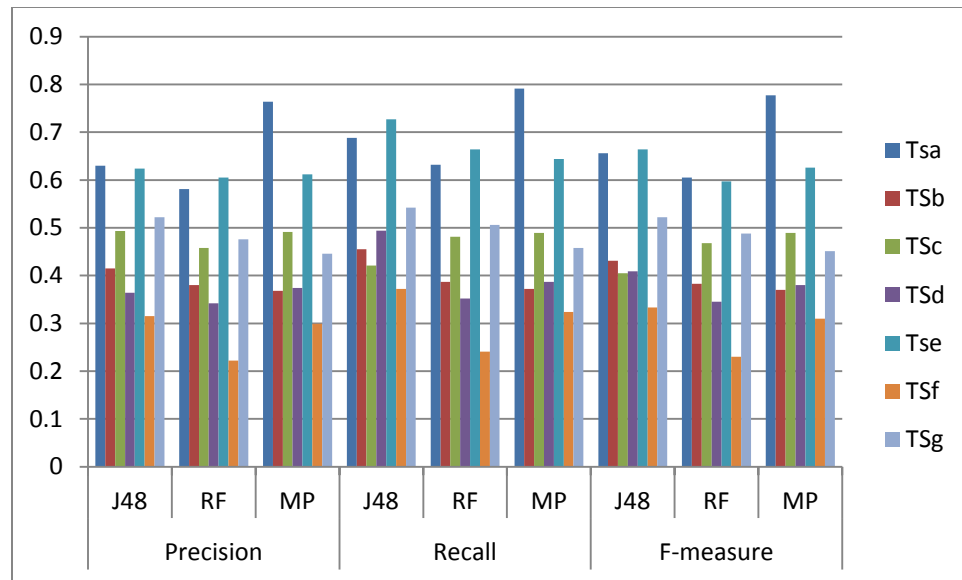


Figure 19: Graph of Table 12

CHAPTER 6: CLUSTERED DATA, CLASSIFICATION STUDY

The clustering process resulted in two classes of viable datasets for mining – the first one represented by three visits datasets, the second represented by four visits datasets, both created from a seed record represented by a patient and used to set the visit distance for a cluster. For three visits, fourteen new datasets were created with attributes listed in the following table. For four visits, five new datasets were created with the same attributes. Total number of datasets analyzed was 1,064.

Table 13: Attributs and Features for the Clustered Dataset

eAttributes	Values of Attributes	Type
Type	Instrument Type	Text
Total Visits	Total Number of Visits	Numeric
Model	Instrument Model	Text
Last_P	Last Patient Type	Text
Instrument	Instrument Name	Text
First_P	First Patient Type	Text
CC	Category of Treatment chosen by Doctor	Text
C	Category of Treatment chosen by Patient	Text
T Difference	Difference in T Score	Numeric
Coefficients	3 coefficients for 3 visits datasets 4 coefficients for 4 visits datasets	Numeric
Angles	3 angles corresponding to visits 1-2, 1-3, and 2-3 (for 3 visits datasets) 6 angles corresponding to visits 1-2, 1-3, 1-4, 2-3, 2-4, and 3-4 (for 4 visits datasets)	Numeric
Sound Features	Sound Level Centroid, Sound Level Spread	Numeric
Recovery Rate	Recovery Rate	Numeric
Text	Stress, Noise, Medical	Boolean
Decision Feature	One of the eight descritized total scores	

In order to test the classifiers, WEKA was used with J48, Random Forest, and the function Multilayer Perceptron (Neural Network) with the eight decision attributes based on the descritized total score. Datasets with the following attributes have been tested:

- 1) Datasets with standard deviations and averages,
- 2) Datasets with coefficients and text,
- 3) Datasets with coefficients and angles,

- 4) Datasets with coefficients only,
- 5) Datasets with angles only,
- 6) Datasets with angles and text,
- 7) Datasets with angles, coefficients and text.

In order to efficiently process the tests and work with the results, a batch file was prepared with carefully named files for processing the ARFF input (descriptive file names). After processing the batch, WEKA output consisted of a dataset of results that can be easily read in Access and analyzed. The attributes stored include the file name, classifier, decision variable (see 5.2 Structure of the Decision Attribute), visits, seed record (for cluster), and then Boolean fields showing the type of features included in the classification including has stats, has coefficients, has angles, and has text. Precision, Recall and F-measure for each test are stored in order to review the accuracy of the classification. Analysis allows matching to the result files for careful analysis.

Our goal is to find and construct new derived attributes yielding possibly the best classifiers for the Tinnitus database. Previously, the top classifier for the unclustered datasets was evidenced by the original Tinnitus dataset with decision feature TSa, Sound Level Centroid, Sound Level Spread, and Recovery Rate features as previously described. The clustering and new features for coefficients and angles improve the classification with the data grouping presenting a more homogeneous dataset. Results are encouraging on the sample datasets; top precision is .884 which represents an improvement over the classification precision of .751 with J48 classification on the original dataset and features Sound Level Centroid, Sound Level Spread and Recovery Rate being present. The new,

improved classification results from WEKA for the clustered dataset appear for TS3 in Figure 20 below.

```
WEKA test with angles, coefficients and text data
File: base_angle_coef_noise_4_d3_[E04-015]_j48.txt
Experiment classifier: J4.8

precision = 0.884
J48 pruned tree
-----

1-4_angle <= -68.749494: a (4.0)
1-4_angle > -68.749494
|   sc_t1_coef <= 32
|   |   total_visits <= 8: b (4.0/1.0)
|   |   total_visits > 8: c (3.0)
|   sc_t1_coef > 32: b (18.0)

Number of Leaves   :    4

Size of the tree   :    7

Time taken to build model: 0.03 seconds
Time taken to test model on training data: 0.01 seconds
```

Figure 20: WEKA Results

The flexibility of the results allows interesting comparisons to be easily made. Figure 21 below shows the comparison of each classification method (J48, Multilayer Perceptron, and Random Forest) and decision variable combination with the maximum precision realized. This particular table does not show the features present when the results are realized after the classification is run.

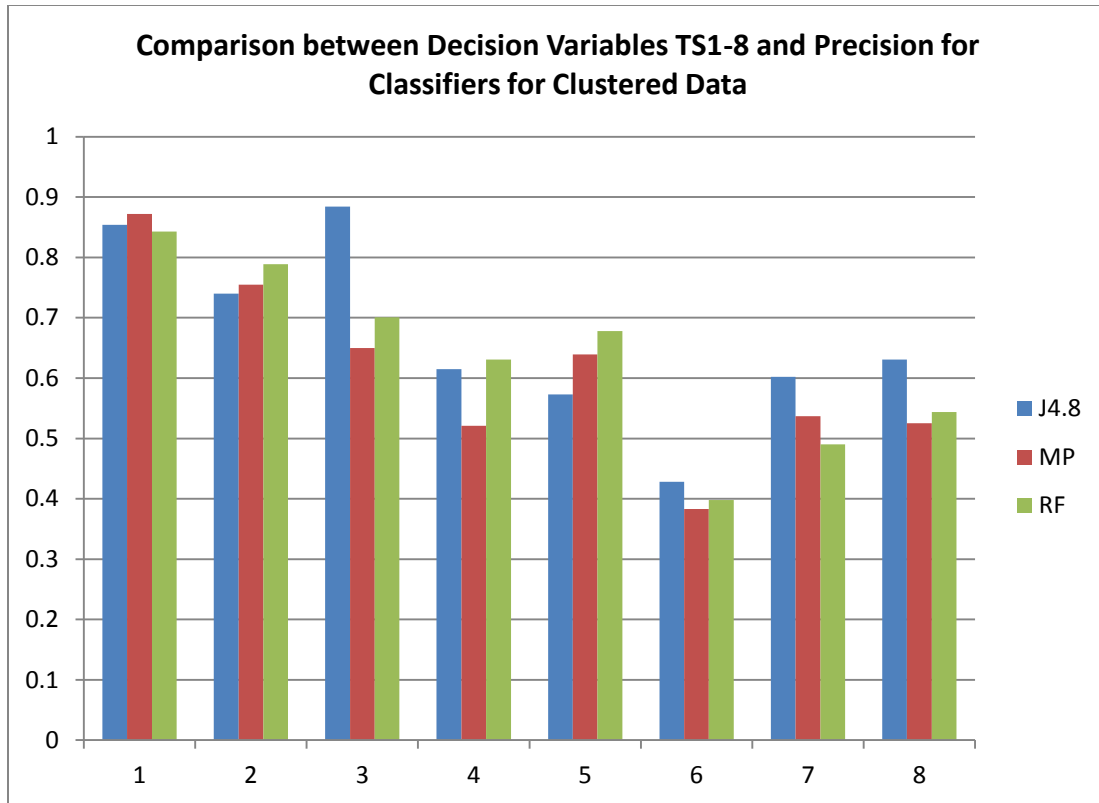


Figure 21: Comparison between Decision Variables 1 through 8

From Figure 14, we see that Decision variable TS3 has the highest precision with J48 classifier. For the clustered dataset, all three decision variables perform approximately the same with decision variable TS1 which is the least demanding on the experiment. TS1 simply splits the Total Score into three components based on whether it is equal to 0, greater than 0, or less than 0.

CHAPTER 7: ACTION RULES

This section mainly concerns the application of action rules to a dataset of new patient visits – each row in a dataset contains information about one patient obtained from the completion of the new Tinnitus Functional Index (TFI). The dataset covers 161 visits represented by 75 unique patients. Only the new patient dataset is used to learn action rules for treatment success based on visits. Of particular interest is the contribution of the new Tinnitus Functional Index and new emotion based features developed from that index. The following topics are covered: Ac4ft-Miner with LISp-Miner for Action Rule Discovery, a new system for action rule discovery called MARDs or Minimal Action Rule Discovery System, the Tinnitus Functional Index and Emotions, the application of the Emotional-Valence Plane to the TFI for new emotions feature development for new patients, and data preparation for action rule discovery.

7.1 Action Rules and Preliminary Research.

An action rule is defined as a rule extracted from an information system that describes a transition that may occur within objects from one state to another with respect to decision attribute that is identified by user, first proposed by Ras and Wieczorkowska [26]. When applied to medical data, action rules show great promise; a doctor can examine the effect of treatment choices on a patient's improved state as measured by an indicator that indicates treatment success, such as the Total Score on the Tinnitus Handicap Inventory.

The attributes used in action rule discovery are identified as stable and flexible with assumption that values of flexible attributes can be changed (for example, a stable attribute in a medical database might be Gender, a flexible attribute might be Hearing Device). The change in flexible attributes can be controlled by the user and used to discover important information about a dataset. For example, action rule discovery can be used to suggest a change on a flexible attribute like hearing device in order to see the changes in treatment effectiveness for tinnitus patients as evidenced by movement to positive total scores from the Tinnitus Handicap Inventory (defined as the decision).

An example of an action rule is as follows (Figure 22): [30]

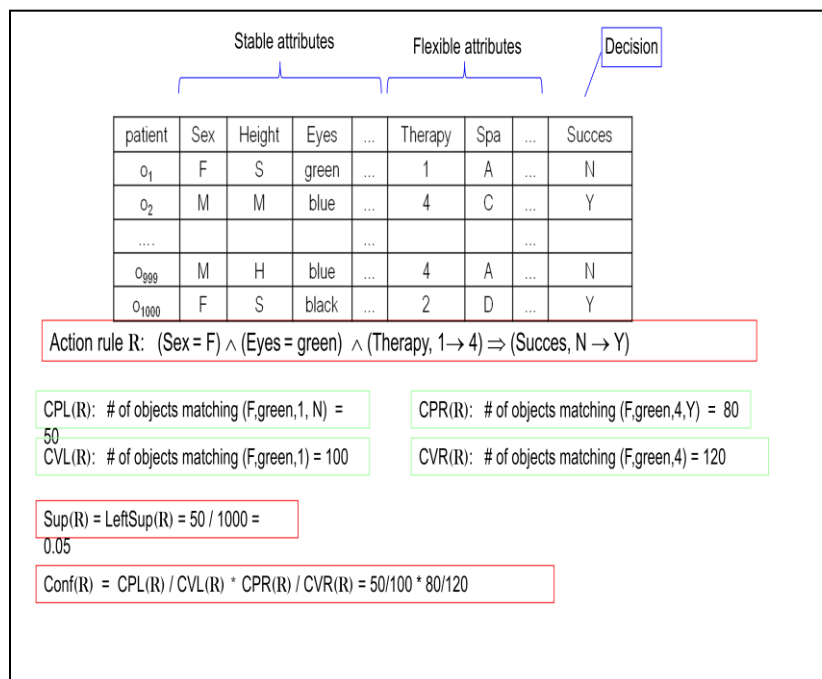


Figure 22 : Example of Action Rule

7.1.1 Preliminary Research with the Clustered Original Database and Action Rule Discovery

In preliminary research with the tinnitus database, the authors applied the flexible temporal features that have been developed and described in Section 4.1 to a decision system with tree classifiers and the action rules construction method previously proposed. Values for numerical attributes in the dataset were hierarchically discretized using a classical method based on entropy or the Gini index. Classification attributes are partitioned into stable and flexible. Before any flexible attributes are used in the process of decision tree construction, all stable attributes must be used first for the action rule construction. This way the decision table is split into a number of decision sub-tables leading to them from the root of the tree by uniquely defined paths built from the stable attributes. Each path defines a header in all action rules extracted from the corresponding sub-table.

The tools used in the preliminary research were SAS with the original tinnitus database in Microsoft Access. WEKA was used for decision tree classification study leading to the construction of Action Rules. A database of 215 patients with at least four visits during the treatment progression with 32 features, including new temporal features, were studied. Data selection was based on a threshold of 2.5 applied to visit, resulting in 14 new datasets with a total of 747 records. See Table 14 for information on the seeds and size of subsets generated per seed.

Table 14. Seeds generated with distance ≤ 3 weeks

Seed Patient	Size of the Subset
03093	62
02038	61
01067	58
04098	57
05024	53
03075	53
05013	52
05011	52
00067	52
04062	51
04008	51
00026	49
01052	48
01038	48
total	747

The above patient records represented by the indicated seeds each have three visits in total in the resulting dataset, with the distance between visits being used to create two features representing distance 1 and distance 2. Input tuples will have three or more visits. This is due to the nature of the data selection algorithm: we may select close visits from the paired patient, only when the paired patient has more visits than the seed patient; therefore, records with small visit numbers tend to collect more similar patterns around them. The features applied in this research (previously presented in section 4.1) include: A1, A2, A3, T1, T2, T3 for the total score of negative emotions; A1, A2, A3, T1, T2, T3 for the total score of functional problems; A1, A2, A3, T1, T2, T3 for the total score of catastrophe; the most important problem in the first visit, the most important problem in the second visit, the most important problem in the third visit, the sound instrument used in the first visit, the sound instrument used in the second visit, the sound instrument used in the third visit, the follow-up method after the first visit, the follow-up method after the second visit, the

follow-up method after the third visit, the dependency on the presence of hyperacusis in the first visit, the dependency on the presence of hyperacusis in the second visit, the dependency on the presence of hyperacusis in the third visit, the real ear measurements in the first visit, the real ear measurements in the second visit, the real ear measurements in the third visit, patient category, the treatment category in the first visit, the treatment category in the second visit, and finally the treatment category in the third visit. The decision attribute is based on whether or not the patient symptoms are improved based on the scores from the Tinnitus Handicap Inventory related to the Scores for Emotions, Function, and Catastrophe and the summed Total Score.

Table 15. Seed generated with distance ≤ 4 weeks

Seed Patient ID	Size of the Subset
03071	47
total	47

Table 15 shows seeds for clustering of the dataset with four visits in total. The features applied in this case include: A3, A4, A5, A6, T3, T4, T5, T6 for the total score of negative emotions; A3, A4, A5, A6, T3, T4, T5, T6 for the total score of functional problems; A3, A4, A5, A6, T3, T4, T5, T6 for the total score of catastrophe; the most important problem in the first visit, the most important problem in the second visit, the most important problem in the third visit, the most important problem in the fourth visit, the sound instrument used in the first visit, the sound instrument used in the second visit, the sound instrument used in the third visit, the sound instrument used in the fourth visit, the follow-up method after the first visit, the follow-up method after the second visit, the follow-up method after the third visit,

the follow-up method after the fourth visit, the dependency on the presence of hyperacusis in the first visit, the dependency on the presence of hyperacusis in the second visit, the dependency on the presence of hyperacusis in the third visit, the dependency on the presence of hyperacusis in the fourth visit, the real ear measurements in the first visit, the real ear measurements in the second visit, the real ear measurements in the third visit, the real ear measurements in the fourth visit, patient category, the treatment category in the first visit, the treatment category in the second visit, the treatment category in the third visit, and the treatment category in the fourth visit. The decision attribute is the same as described for the previous clustered dataset.

Decision tree study was performed using J48 in WEKA, a system previously described. The evaluation was for positive recovery from functional problems, negative emotions, and catastrophe. All classifiers were 10-fold cross validation with a split of 90% training and 10% testing.

In this initial research on action rules, several interesting rules we discovered are listed below:

Rule 1. generated from seed 02038: (original tinnitus database, clustered): [(patient category=2) \wedge (A1 of total score ≤ 3.7) \wedge (initial real ear measurements, $y \rightarrow n$)] \Rightarrow (positive recovery of catastrophe, $n \rightarrow y$)

Support: 4, Confidence: 75.3%

The meaning of the above rule is as follows: if patients indicate hearing loss as a significant subjective problem and tinnitus as a significant problem, and also have A1 of the total score less than 3.7, having real ear measurements in the first visit or not decides if they will have improvements in terms of catastrophe scores after Tinnitus Retraining Therapy.

Rule 2. generated from seed 02038 (original tinnitus database, clustered) : [(patient category=3) \wedge (A1 of total score \leq 3.7) \wedge (follow up method, “counseling” \rightarrow “telephone”)
 \Rightarrow (positive recovery of catastrophe, $n \rightarrow y$)]

Support: 2, confidence: 92.3%

The meaning of the above rule is as follows: if patients have a primary problem of hyperacusis and are treated for this condition with a specific Tinnitus Retraining protocol that involves use of wearable sound generators or combination instruments, and have A1 of the total score less than 3.7, the change of follow up method from counseling to telephone indicates improvements in terms of catastrophe scores after treatment with Tinnitus Retraining Therapy. Additionally, the rule strongly suggests that method “telephone” in the mentioned condition means improvements in the catastrophe score, where this side of the action rule has a support of 13.

Rule 3. generated from seed 04062 (original tinnitus database, clustered): [(patient category=3) \wedge (A1 of total score \leq 1.1) \wedge (initial dp=n) \Rightarrow (positive recovery of negative emotion=y)]

Support: 9, confidence: 69.2%

The meaning of the above rule is as follows: if a patient has a primary problem of hyperacusis and is treated for this condition with a specific Tinnitus Retraining protocol that involves use of wearable sound generators or combination instruments, and has A1 of the total score less than 1.1, and there is no dependency on the presence of hyperacusis, he or she may have improvements in terms of their score related to negative emotions after Tinnitus Retraining Therapy.

Rule 4. generated from seed 04062 (original tinnitus database, clustered): [(patient category=4) \wedge (A1 of total score ≤ 1.1) \wedge (T1 of total score, $\leq 2 \rightarrow > 2$) \Rightarrow (positive recovery of negative emotion, $n \rightarrow y$)]

Support: 4, confidence: 66.7%

The meaning of the above rule is as follows: if patients are relatively uncommon and suffer from a condition in which their tinnitus or their hyperacusis is significantly worsened because of exposure to certain types of sounds, and if their A1 of the total score is not greater than 1.1 and their T1 of the score of catastrophe changes from less than or equal to -2 to greater than 2, then they may begin to have improvements on negative emotions. More so, when the T1 of the score of catastrophe is greater than 2, the mentioned typed of patients will always have improvement on negative emotions, as this side of the mentioned action rule has support of 7 and confidence of 100%.

Rule 5. generated from seed 02038 (original tinnitus database, clustered): [(patient category=3) \wedge (A1 of total score ≤ 1.1) \wedge (T2 of total score ≤ 12) \wedge (most important problem, $H \rightarrow (T \mid L)$)] \Rightarrow (positive recovery of negative emotion, $n \rightarrow y$)

Support: 7, confidence: 63.2%

The meaning of the above rule is as follows: if a patient has the primary problem of hyperacusis and is treated for this condition with a specific TRT protocol that involves use of wearable sound generators or combination instruments, and has A1 of the total score less than 1.1 and T2 of the total score less than or equal to 12, the change of the most importance problem from hyperacusis to tinnitus or hearing loss indicates improvement on negative emotions.

Rule 6. generated from seed 00026 (original tinnitus database, clustered): $[(T1 \text{ of total score} \leq 2) \wedge (A2 \text{ of total score} \leq 12) \wedge (\text{the most important problem in the last visit is "Tinnitus"}) \wedge (\text{rem of the second visit, } y \rightarrow n)] \Rightarrow (\text{positive recovery of negative emotion, } n \rightarrow y)$

Support: 8, confidence: 66.7%

The meaning of the above rule is as follows: if a patient has T1 of the total score not great than 2 and T2 of the total score not greater than 12, and if tinnitus is the most important problem in the last visit, stopping the real ear measurements in the second visit means improvement of the negative emotions.

Rule 7. generated from seed 00026 (original tinnitus database, clustered): $[(T1 \text{ of total score} > -4) \wedge (T2 \text{ of total score} \leq -2) \wedge (T2 \text{ of catastrophe} > -4) \wedge (p2 = T) \vee (p2 = L)] \Rightarrow (ta_sc_f13 \leq 0)$

Support: 16, confidence: 76.2%

The meaning of the above rule is as follows: if a patient has T1 of the total score greater than -4 and T2 of the total score not greater than -2 and T2 of the catastrophe greater than -4 and the most important problem of the second visit is either tinnitus or sound loss, then he or she will have improvement in terms of functional problems.

7.1.2 Summary of Preliminary Research on Original Database and Action Rules

Preliminary research on action rules showed much promise toward leading to the discovery of new and interesting rules for tinnitus decision support based on clustering the dataset according to three and four visit sets and generating new temporal features.

Improved action rule discovery engines are explored in continuing research, along with some new and exciting temporal and emotions based treatment features, leading to relevant

results presented in Chapter 8. One additional, more sophisticated tool and one new tool for action rule discovery will be explored in the continuing sections, along with some interesting and useful rules discovered from the Tinnitus database.

7.2 LISp-Miner for Action Rule Discovery.

LISp-Miner with the 4ft-Miner procedure is a part of the robust LISp-Miner system developed by Dr. Jan Rauch and his colleagues (<http://lispminer.vse.cz> and [23] [24]).

LISp-Miner includes an advanced system of software modules that have been developed to implement classification and action rule discovery algorithms on data sets. The 4ft-Miner procedure is used in this research to discover new action rules in the tinnitus datasets with respect to new patients (those completing the Tinnitus Functional Index).

7.2.1 Background.

LISp-Miner takes an approach to the construction of action rules based on the GUHA method and its implementation [23] [24]. The action rules in LISp-Miner are called “G-action rules”.

7.2.2 GUHA and LISp-Miner.

GUHA is realized by GUHA-procedures, and has been in use since the 1960’s as a method of exploring data. To summarize, input to a GUHA system consists of a dataset and meta-data which describes patterns which are of interest in the data. In essence, GUHA as implemented in LISp-Miner mines for association rules.

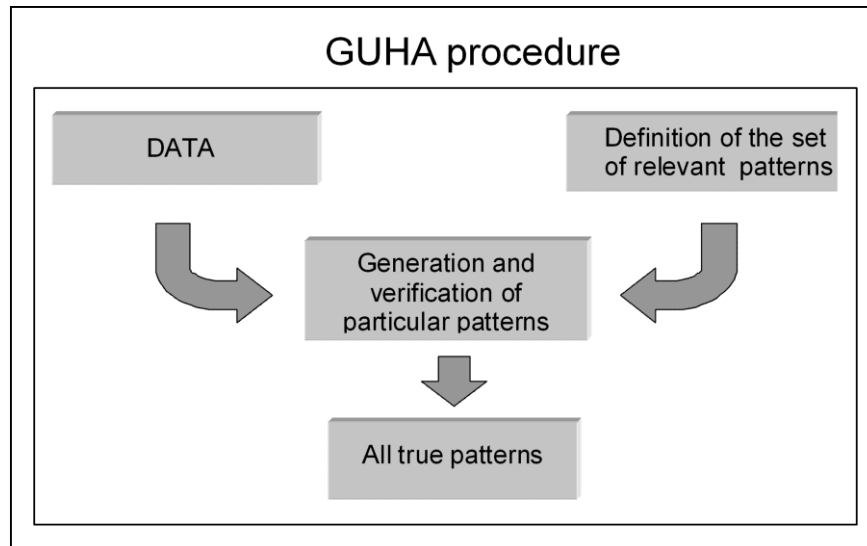


Figure 23: GUHA Procedure

LISp-Miner mines for *all* true patterns in the data, limited only by the size of the dataset and the definition of the relevant patterns of interest to the user. For example, if you have 20 attributes of interest and you want to include six of these to form patterns in mining, your number of relevant rules that can include combinations of these patterns can be incredibly large. LISp-Miner allows some reduction of the patterns of interest but this requires specific knowledge of the ontology that the dataset satisfies. You can further limit the values in the attributes that are being mined by adding left and right cuts, effectively reducing the values of interest for specific variables. Additionally, variables are defined as stable and flexible with respect to the decision variable of interest [23] [24].

7.2.3 Association Rules and LISp-Miner.

LISp-Miner documentation (found at <http://lispminer.vse.cz>) describes an association rule as “. . .commonly understood to be an expression of the form $X \rightarrow Y$, where X and Y are sets of items. The association rule $X \rightarrow Y$ means that transactions containing items of set X tend to contain items of set Y . There are two measures of intensity of an association rule – confidence and support.” [23] [24] [25].

In association rule discovery, the goal is to find all association rules of the following form: $X \rightarrow Y$. The desire is that the support and confidence are higher than user set thresholds for the level of minimum confidence and minimum support. [23] [24] [25]. Confidence or accuracy is the proportion of examples predicted accurately, expressed as a proportion of all examples that apply. Support, also called coverage, is the actual number of examples or instances predicted correctly. Confidence can be as high as 100% [28].

Association rules are similar to classification rules but have the added ability to predict any attributes and combinations of attributes. Many association rules can be generated from a set of data, therefore, it is necessary to specify desired thresholds for minimum confidence and support. The coverage of an association rule then becomes the number of instances that the rule predicts correctly, with the confidence being the number of instances predicted correctly divided by the number of instances that actually apply to the rule. (WEKA p. 69). In the first step all frequent itemsets are found (set of items meeting minimum support); the second step generates those that meet the minimum confidence.

The procedure in LISp-Miner called Ac4ft-Miner mines for association rules of the form $\Phi \approx \Psi$ with Φ and Ψ representing Boolean attributes antecedent and succedent respectively. The association rule represented by $\Phi \approx \Psi$ means that the antecedent and

succedent are associated in a way represented by \approx which is called the “4ft-quantifier”.

This is represented by a quadruple data matrix represented as:

M	ψ	$\neg \psi$
φ	a	b
$\neg \varphi$	c	d

Figure 24: Ac-4ft Quantifier

The Ac4ft-Miner procedure can be best understood as providing an enhancement to the action rules mining procedures. The a-priori algorithm of association rules discovery is not employed, and the procedure that is used follows a complex bit-string method; an explanation is provided from the important work by Rauch and Simunek:

“We assume that the attribute A has k particular values a_1, \dots, a_k . The expression $A(a_1)$ denotes the Boolean attribute that is true if the value of attribute A is a_1 etc. . . . This approach is based on representing each possible value of an attribute by a single string of bits. In this way it is possible to mine for association rules of the form, for example, $A(_) \wedge B(_) \rightarrow C(_)$ where $(_)$ is not a single value but a subset of all the possible values of the attribute A. The expression $A(_)$ denotes the Boolean attribute that is true for a particular row of data matrix if the value of A in this row belongs to $(_)$, and the same is true for $B(_)$ and $C(_)$. The bit string approach means that it is easy to compute all the necessary frequencies. Then we can mine not only for association rules based on confidence and support but also for rules corresponding to various additional relations of Boolean attributes including relations described by statistical hypotheses tests. . . . The bit string approach also makes it easy to mine for conditional association rules that are mentioned . . .” [23] [24] [25]

The utilization of the bit string method has been documented as a successful technique for rule discovery (Simunek, Academic KDD Project LISp-Miner –see 13). The complexity of this method leads to an algorithm that produces a maximum number of rules with large numbers of attributes involved in the rules. This is probably more valuable for the medical researcher in the long run, but the problem is the system complexity associated

with rule discovery and the knowledge of the ontology in order to establish the necessary system parameters for rule discovery and rule interpretation.

The disadvantages of Arc4ft-Miner in LISp Miner are improved with the implementation of the new Action Rule Mining Engine, MARDs.

Application of the Ac4ft-Miner System to the Tinnitus database in LISp-Miner along with results is discussed in Section 8 of this work. Of importance is the complexity associated with action rule discovery with Ac4ft-Miner as compared to the results from MARDs discussed in the next section.

7.3 A New Application for Action Rule Discovery

This section presents a new system for Action Rule discovery called Minimal Action Rule Discovery system or MARDs. The word Minimal is indicative of the reduced time of mining for action rules by the system and the simplicity and minimal length of the rules as compared to the extensive, yet complex rules discovered by Arc-4ftMiner with LISp. MARDs is developed in C++ and has the ability to quickly generate the shortest action rules (involving a maximally reduced number of relevant attributes). The MARDs system also allows the research to generate important knowledge that will facilitate more extensive analysis using a system like Arc4ft-Miner.

The goal of MARDs is to generate the smallest possible subset of relevant action rules. The system generates frequent item sets and then compares these to the thresholds of support and confidence that are imposed by the user.

For example, if a rule is generated as $a \rightarrow b$ and the rule is under the minimum thresholds for support and confidence, this will be considered as the minimal rule and no further rules

of the form $ca \rightarrow b$ will be generated (classification part of $a \rightarrow b$ will be not extended). In other words, MARDs generates rules of the shortest length.

Obviously, this system is better in terms of time complexity than Arc4ft-Miner which may generate all the rules. A disadvantage is that the user does not gain the extensive knowledge into the generated rules that may be provided by the inclusion of more attributes representing the stable and flexible features affecting the decision attribute. If the research involves a medical database like the tinnitus database, it is assumed that the physician would be more interested in the detailed yet time-costly rules.

Even with this disadvantage, however, MARDs is quite important to the field of medical data mining and rule discovery. The data mining expert is most likely not the expert in the ontology related to the investigated medical domain; MARDs mining provides valuable insight into a dataset at a minimal cost. This insight can make the researcher more effective as they implement more complex and extensive mining methods such as those represented in LISp. When you are learning the problem that is of interest, less costly and shorter rules might be extremely valuable.

7.4 The Tinnitus Functional Index and Emotions.

The new Tinnitus Functional Index has been previously presented in this research. The enhanced capability that the TFI presents for measuring the patient emotional state is one of the interests of Action Rule discovery.

Much research has been performed on the role that the auditory system plays on tinnitus. Tinnitus perception is generally considered not to be pathologic as 94% of individuals without prior tinnitus will realize the condition when requested in a sound proof chamber for a short period of time. Dr. Jastreboff has based Tinnitus Retraining Therapy on the effort to reduce the discomfort and annoyance associated with tinnitus through counseling. Patients report strong emotional reactions to tinnitus, indicating the involvement of the limbic and autonomic nervous systems; tinnitus retraining therapy through counseling works to improve the emotional reaction to tinnitus through counseling.

As a patient continues Tinnitus Retraining Therapy, they complete a Tinnitus Handicap Inventory during each visit. The THI was previously presented and a section of the inventory relates to a score that patients receive for emotions. The total score combines the questions related to emotions, patient function, and the catastrophic nature of tinnitus and improvements in tinnitus can be measured by a lowering of the total score. The questionnaire is rated by a 4 representing yes, a 2 representing sometimes, and a 0 representing no to questions that are targeted toward the effect tinnitus has on the life of the patient.

The Tinnitus Functional Index was developed by a group of researchers in partial response to the need to develop improved measures for assessing ongoing treatment as compared to measuring and screening patients during intake. [29]. The new Tinnitus

Functional Index is administered to patients designated as new in the second dataset received from Dr. Jastreboff; this represents 161 visit tuples for 75 unique patients. The Index uses an eleven point scale as previously presented. The relatively coarse response scale for the THI (3 levels, Yes=4, Sometimes=2, No=0) is considered to be less sensitive to the effect of treatment than the eleven point scale for the TFI [20]. Both questionnaires are administered in conjunction with almost every treatment visit for the new patients. One of the goals of the new TFI questionnaire is to show improved responsiveness to changes in patients, including emotional based change, based on the treatment progress over time. [29]

7.5 Emotions Feature Development.

The new features for emotion developed for this study are described in Section 3.4, The features E1, E2, E3, and E4 along with the Total Score from the Tinnitus Handicap Inventory were applied to patient visit tuples. Additionally, features related to the improvement in scores were calculated as numeric differences and Boolean attribute + or – showing improvement or negative improvement in the patient looking forward to the next visit. In this way, a patient tuple shows the treatment that the patient received for a particular visit and the THI and TFI scores and Emotional values for the next visit showing treatment factor success relative to the treatment received for a particular visit. Construction of the individual records for the patient in this manner creates multiple tuples for each patient related to total number of visits – 1. Association and action rule mining can effectively occur, yielding new and interesting rules related to the new features for emotions and treatment effectiveness (Table 2: Tinnitus Functional Index).

7.6 Data Preparation for Action Rule Discovery.

The data sample for our experiments is not as large as would be desired for more significant results; yet the application of the sample to action rule discovery with 4ft-Miner and the new action rule discovery engine is deemed applicable for indications for rules for a tinnitus decision support system.

Significant data preparation occurred before the two action rule discovery systems could be applied:

- 1) New patients were identified as those completing the Tinnitus Functional Index from the new dataset received from Dr. Jastreboff late in the research. Patients with one visit were eliminated as one visit did not give the information related to treatment success, determined by measurements on subsequent visits.
- 2) A subset of the dataset containing attributes and new features (previously explained) was prepared by visit.
- 3) Boolean and numeric features were added showing treatment success for a particular visit, based on scores in the THI and TFI for the next visit for the patient. The last visit for each patient was removed as there were no indicators that could be used to show treatment success after the last visit.

For Arc4ft-Miner in LISP-Miner, antecedent and succedent attributes were identified along with relevant attributes and other important information (including partition which attributes are stable and flexible). For the new MARDs Action Rule engine, stable and flexible variables were identified. Minimum confidence and support levels were established for both systems. Several iterations of action rule discovery then occurred in order to learn the capabilities of each system and to discover interesting and useful rules based on new features, specifically features tied to emotions.

Results of the experiments are presented in Chapter 8.

CHAPTER 8: ACTION RULES – EXPERIMENT AND RESULTS

Action Rule mining was accomplished using LISp-Miner's Ac4ft-Miner and MARDs - the new Action Rule discovery engine. Results are presented along with a comparison of the application of the two systems for task.

8.1 Action Rules Ac4ft-Miner with LISp-Miner.

Preparation of the tinnitus database involved developing the meta-data and data for 142 columns (107 total new tuples) including a unique ID and unique Patient-ID as identifiers for each patient. Other attributes are identified in Appendix A. Important to this research are the new attributes for Emotion developed from the Tinnitus Functional Index, new to the patients for this study on the extended database. Additionally, attributes were developed to show the change in the treatment or improvement in the patient by Boolean feature (+ or -) and by numerical change for numerous columns; these new features are important for action rule discovery and are included in the detailed listing of features in the Appendix. A summary of the attributes and features is found in Table 16 and will be used to simplify the discussion of the mining tasks.

Table 16: Attributes and Features used in LISp-Miner and Arc4ft-Miner

Abbreviation	Characteristics	List of attributes
Initial state		
BASIC	Patient's basic characteristics	<i>ProblemTHL, Misophonia, Sc_T</i>
TRT	Patient's initial state – questions from TRT	H_Sv, H_An, H_EL, H_pr, HI_pr, Aw%T, An%T, Tch, T_Sv, T_An, T_EL
QQQ	Patient's initial state – Tinnitus Function Index	Q1, ..., Q25
E_SCORE	Patient's initial state – emotion score	E1_SCORE_TFI, E2_SCORE_TFI, E1_SCORE_TFI, E4_SCORE_TFI
Treatment		
TRTM	Treatment	<i>Instrument, Trtmnt_Cat_Patient, Trtmnt_Cat_Dr</i>
Results of treatment		
IMPR_TRT	Improvements in attributes related to the TRT	Impr_in_H_Sv, Impr_in_H_An, Impr_in_H_EL Impr_in_H_pr, Impr_in_HI_pr, Impr_in_Aw%T Impr_in_An%T, Impr_in_Tch, Impr_in_T_Sv Impr_in_T_An, Impr_in_T_EL,
CHG_E	Changes in emotional score	CHG_IN_E1, CHG_IN_E2, CHG_IN_E3, CHG_IN_E4, CHG_IN_Q1,

The LISp-Miner with AC4ft-Miner includes useful features to examine the data, meta-data, and value frequencies in the AC4ftTask module, implemented with the system.

8.1.1 Input and Task Identification

The input for Ac4ft-Miner in LISp-miner consisted of a data matrix representing the prepared data and meta-data associated with attributes of interest in the tinnitus datasets, as described above. Mining tasks of interest are given in Table 17. Antecedents can be identified as stable or flexible, and can be further refined. Succedents also can be identified as stable or flexible, with conditions and cuts optional on the data. Antecedent, succedent and condition together are called cedents [30]. Simply stated, the antecedent can be one or

more attributes or features on the left hand side of the rule, and succedents are on the right hand side.

Table 17: Mining Tasks of Interest

Task	Antecedent		Succedent	
	stable	flexible	stable	flexible
Test	E1_Score	Instrument	not used	An%T
T_01	BASIC	TRTM	not used	IMPR_TRT
T_02	BASIC	TRTM	not used	CHG_E
T_03	BASIC, TRT	TRTM	not used	IMPR_TRT
T_04	BASIC, TRT	TRTM	not used	CHG_E
T_05	BASIC, QQQ	TRTM	not used	IMPR_TRT
T_06	BASIC	E_SCORE	not used	IMPR_TRT

8.1.2 Preliminary Rule Discovery and Discussion of Resulting Output

Rules were first discovered for a small subset of the data (Task T_01 above); the analytical question of interest is “what is the effect of changing instrument for a particular level of E1_Score on the Annoyance of Tinnitus?”. This, in essence, will give hints to the effect that emotions have on Annoyance of Tinnitus. Of interest is the change in score E1 representing “Energetic Positive” (sum of Questions 3 in control, 5 cope, and 20 enjoyment of life, each on a scale of 0 to 10 from the Tinnitus Functional Index with 0 being a positive score value) and the change in Improvement in the Annoyance of Tinnitus as presented by + representing improvement from one visit to the next (in other words, the treatment reflected on the particular tuple shows improvement as measured by looking ahead to the next data value in that category).

Domain knowledge is necessary to implement this accurately. The E1 score is a value from 0 to 30 (representing the summed values of three questions related to “Energetic Positive” each on a scale from 0 to 10). An examination of the frequencies in the category is

necessary in order to determine meaningful cuts for the mining software. The category frequencies are in Table 18 for feature E1.

Table 18: Attribute categories frequency analysis for feature E1
(possible value 0 to 30)

#	E1_score_tfi	Freq %	Frequency	Cummul. Freq %	Cum. Freq.
1	0	2.8 %	3	2.8 %	3
2	(0;3>	1.9 %	2	4.7 %	5
3	(3;6>	1.9 %	2	6.5 %	7
4	(6;9>	17.8 %	19	24.3 %	26
5	(9;12>	19.6 %	21	43.9 %	47
6	(12;15>	8.4 %	9	52.3 %	56
7	(15;18>	13.1 %	14	65.4 %	70
8	(18;21>	7.5 %	8	72.9 %	78
9	(21;24>	18.7 %	20	91.6 %	98
10	(24;27>	5.6 %	6	97.2 %	104
11	(27;30>	2.8 %	3	100.0 %	107

The mining software allows cuts to be placed in the numeric ranges for the scores and cuts of interest were placed on 6:9 (left cut) and 15:18 (right cut) for groupings of values of E1. The Antecedent (stable variable) was identified as the E1 score from the Tinnitus Functional Index based on the identified cuts (stable) and the instrument type was identified as a flexible antecedent. The succedent (decision or right hand side) was identified as the Improvement in the Annoyance of Tinnitus represented by the feature Impr_In_An%T with values + representing Improvement and – representing lack of improvement. No conditions were identified as a part of the cedents. One of the rules from the results will be described in order to define items of interest in the output and to serve as a base for the remaining discussion of the mining with LISp. The output for the rule is found in Table 19:

Table 19: Hypothesis for Resulting Rule				
State Before: E1_score (6:9> . . (15:18>) && Instrument (GH) *** Impr in An%T (-)				
State After: E1_score (6:9> . . (15:18>) && Instrument (GOTE) *** Impr in An%T (+)				
	Succedent	¬ Succedent	Succedent	¬ Succedent
Antecedent	9	0	9	20
¬ Antecedent	58	40	31	47

The resulting rule can be stated as for E1 scores with instrument GH and Improvement in the Annoyance % of Tinnitus showing a lack of improvement, if the Instrument is changed to instrument GOTE then improvement in the Annoyance % of Tinnitus goes from – to + or positive. Obviously, of interest is the support and confidence of the stated rule. Before and after states are given in the results. In the before state, there are 9 patients with E1 score between 6:9 and Instrument GH with Improvement in Tinnitus (-).

The confidence of the before state is $9 / [9 + 0]$ or 100% with support 9. The state after has 9 patients out of 20 with a change in instrument to GOTE and Improvement in Annoyance of Tinnitus to positive (a good change). The confidence is $9 / [9 + 20]$ or 31% with a support of 9.

This preliminary result has low confidence but did show promise with respect to the association between the emotional scores and the Annoyance of Tinnitus. Analysis of results will continue in tabular form (related to Tasks identified above) and will include the question of interest, the input (antecedent and succedent, stable and flexible), conditions and cuts, the output including the number of hypotheses (rules) found and interpretation of rules deemed useful to our research, and comments.

8.2 Analytical Questions and Rules from LISp-miner

In this section, Tasks 1 through 6 detailed in Table 17 above will be presented. For each task, the analytical question will be presented, the input parameters consisting of the stable and flexible parts of the antecedent and succedent for each question, and the output including system cost and the before/after grid with support and confidence calculations will be shown. For each task, many rules are generated and the rules presented will be rules of interest.

8.2.1 Task 01.

Task 01 is a rule that specifies if the Total Score from the Tinnitus Handicap Inventory is in the mild range and the Instrument is GH, if the Instrument is changed to GOTE then improvement in Tinnitus moves from – to + with a confidence of .47. From this set of 94 rules generated from the Task 01 hypothesis, many showed that improvement in tinnitus would occur if the instrument changed from GH to GOTE.

Table 20: TASK 01				
Analytical Question: What treatments cause an improvement in tinnitus as measured by attributes and features in IMPR_TRT?				
INPUT				
	Antecedent		Succedent	
Stable Part	BASIC		Not Used	
Variable Part	TRTM		IMPR_TRT	
OUTPUT				
Number of rules found:	94	Number of verifications:	13488	
Duration (PC dependent):		0h 0m 17s		
Analysis of Rule of Interest (No. 69, ID 75)				
Antecedent: Sc_T(mild): (Instrument(GH) -> Instrument (GOTE))				
Succedent: (Impr_in_T_AN(-) -> Impr_in_T-AN(+))				
Condition: (empty)				
State Before: Sc_T (mild) && Instrument (GH) Impr_in_T_AN(-)				
State After: Sc_T (slight, mild) && Instrument (GOTE) Impr_in_T_AN(+)				
RESULT GRID (Before and After state)				
	Succedent	¬ Succedent	Succedent	¬ Succedent
Antecedent	9	0	8	9
Support and Confidence				
Support:	Before: .08 (n=107), After: .07 (n=107)			
Confidence:	Before: 1 (9/(9+0)), After: .47 (8/(8+9))			

8.2.2 Task 02.

Table 21: TASK 02				
Analytical Question: What treatments cause change in emotion scores as measured by attributes and features in CHG_E?				
INPUT				
	Antecedent		Succedent	
Stable Part	BASIC		Not Used	
Variable Part	TRTM		CHG_E	
OUTPUT				
Number of rules found:	35	Number of verifications:	18180	
Duration (PC dependent):		0h 0m 15s		
Analysis of Rule of Interest (No. 24, ID 23)				
Antecedent: Problem_THL (T_First): (Instrument(GHI) -> Instrument (GH))				
Succedent: (Chg_in_E4>0) -> (Chg_in_E4(0))				
Condition: (empty)				
State Before: Problem_THL (T_First) &&(Instrument(GHI) (Chg_in_E4>0)				
State After: Problem_THL (T_First) &&(Instrument(GH) (Chg_in_E4(0))				
RESULT GRID (Before and After state)				
	Succedent	¬ Succedent	Succedent	¬ Succedent
Antecedent	5	0	5	6
Support and Confidence				
Support:	Before: .05 (n=107), After: .05 (n=107)			
Confidence:	Before: 1 (5/(5+0)), After: .45 (5/(5+6))			

Task 02 is a rule that specifies if the Total Score from the Tinnitus Handicap Inventory is in the mild range and the Instrument is GH, if the Instrument is changed to GHI then the change in new feature E4 moves from >0 to 0 and greater with a confidence of .47. The feature CHG_in_E4 measures the change in the E4 score based on a treatment and looking ahead to the E4 score after the treatment for a specific visit. If a patient is getting better, the change in E4 should go from a higher number to a lower number. This rule needs to be carefully examined; the instrument change to GH from GHI corresponds to a worsening in the change in emotions score. Many rules focus on the effect that a change in instrument has on the patient in terms of Total Score and Emotions; this is significant.

8.2.3 Task 03.

Task 03 is a rule that specifies if the Total Score from the Tinnitus Handicap Inventory is in the slight or mild range and the Tinnitus Severity is in the interval $(2,3> \dots 5,6)$ and Instrument is GH, if the Instrument is changed to GOTE then improvement in the Annoyance of Tinnitus moves from $-$ to $+$ with a confidence of .47.

Table 22: TASK 03				
Analytical Question: What treatments cause an improvement in patient scores as measured by attributes and features in TRT (Initial patient questionnaire values)?				
INPUT				
	Antecedent		Succedent	
Stable Part	BASIC, TRT		Not Used	
Variable Part	TRTM		IMPR_TRT	
OUTPUT				
Number of rules found:	118	Number of verifications:	1002288	
Duration (PC dependent):		0h 13m 39s		
Analysis of Rule of Interest (No. 75, ID 89)				
Antecedent: Sc_T(slight, mild) & T_Sv((2;3>...(5;6>): (Instrument(GHI) -> Instrument (GOTE))				
Succedent: (Impr_in_T_An(-) -> Impr_in_T_An(+))				
Condition: (empty)				
State Before: Sc_T(slight, mild) & T_Sv((2;3>...(5;6>) && (Instrument(GH) Impr_in_T_An(-)				
State After: Sc_T(slight, mild) & T_Sv((2;3>...(5;6>) && (Instrument(GOTE) Impr_in_T_An(+)				
RESULT GRID (Before and After state)				
	Succedent	¬ Succedent	Succedent	¬ Succedent
Antecedent	9	0	9	10
Support and Confidence				
Support:	Before: .08 (n=107), After: .08 (n=107)			
Confidence:	Before: 1 (9/(9+0)), After: .47 (9/(9+10))			

8.2.4 Task 04.

Table 23: TASK 04				
Analytical Question: What treatments cause an change in patient scores as measured by attributes and features in Emotion Scores (Tinnitus Functional Index)?				
INPUT				
	Antecedent	Succedent		
Stable Part	BASIC, TRT	Not Used		
Variable Part	TRTM	CHG_E		
OUTPUT				
Number of rules found:	>199	Number of verifications:	1252860	
Duration (PC dependent):		0h 10m 11s		
Analysis of Rule of Interest (No. 3, ID 3)				
Antecedent: Sc_T(slight, mild) & T_Sv((1;2>...(4;5>): (Trtmnt_Cat_Dr(2) -> Trtmnt_Cat_Dr(3))				
Succedent: (Chg_inE1(>0) -> Chg_in_E1(0))				
Condition: (empty)				
State Before: Sc_T(slight, mild) & T_Sv((1;2>...(4;5>) && (Trtmnt_Cat_Dr(2) Chg_in_E1(>0)				
State After: Sc_T(slight, mild) & T_Sv((1;2>...(4;5>) && (Trtmnt_Cat_Dr(3) Chg_in_E1(0)				
RESULT GRID (Before and After state)				
	Succedent	¬ Succedent	Succedent	¬ Succedent
Antecedent	7	0	7	7
Support and Confidence				
Support:	Before: .07 (n=107), After: .07 (n=107)			
Confidence:	Before: 1 (7/(7+0)), After: .5 (7/(7+7))			

Task 04 provides some interesting results related to patients being treated with mild tinnitus. When the doctor changes the treatment category for the patient from 2 to 3, this is a change in category that represents a move from the patient being categorized with “tinnitus significant and hearing loss” to “tinnitus irrelevant and hyperacusis present”. The rule shows that as this change is made, the patients emotions do not improve based on the E1 score which represents emotions related to the patient’s ability to be in control, cope, and enjoy life. A higher CHG_in_E1 score is better and the rule shows that the removal of tinnitus in the categorization made by the doctor somehow negatively affects the emotions in this category.

8.2.5 Task 05

Table 24: TASK 05				
Analytical Question: What treatments cause an improvement in patient scores as measured by attributes and features in TRT (Initial patient questinnaire values)?				
INPUT				
	Antecedent		Succedent	
Stable Part	BASIC, QQQ		Not Used	
Variable Part	TRTM		IMPR_TRT	
OUTPUT				
Number of rules found:	79	Number of verifications:	574704	
Duration (PC dependent):		0h 6m 18s		
Analysis of Rule of Interest (No. 73, ID 64)				
Antecedent: Sc_T(mild, moderate) & Q21(<=(1;2): (Instrument(GH) -> (Instrument(GOTE))				
Succedent: (Impr_in_T_pr(-) -> Impr_in_T_pr(+))				
Condition: (empty)				
State Before: Sc_T(mild, moderate) & Q21(<=(1;2)&&Instrument(GH) Impr_in_T_pr(-)				
State After: Sc_T(mild, moderate) & Q21(<=(1;2)&&Instrument(GOTE) Impr_in_T_pr(+)				
RESULT GRID (Before and After state)				
	Succedent	¬ Succedent	Succedent	¬ Succedent
Antecedent	7	0	7	7
Support and Confidence				
Support:	Before: .07 (n=107), After: .07 (n=107)			
Confidence:	Before: 1 (7/(7+0)), After: .5 (7/(7+7))			

The meaning of the above is that the Total Score from the Tinnitus Handicap Inventory is in the mild to moderate range and Question 21 (tinnitus effect on relationships) is in category 2 or less (cumulative frequency for Q21 at this value is 45.8%), if the instrument is changed from GH to GOTE then Improvement in Tinnitus as a problem is realized.

Categories for Question 21 are found in Table 25 below.

Table 25: Categories for Question 21

#	Q21	Freq %	Frequency	Cummul. Freq %	Cummul. Frequency
1	(-1;0>	20.6 %	22	20.6 %	22
2	(0;1>	12.1 %	13	32.7 %	35
3	(1;2>	13.1 %	14	45.8 %	49
4	(2;3>	11.2 %	12	57.0 %	61
5	(3;4>	0.9 %	1	57.9 %	62
6	(4;5>	14.0 %	15	72.0 %	77
7	(5;6>	6.5 %	7	78.5 %	84
8	(6;7>	10.3 %	11	88.8 %	95
9	(7;8>	5.6 %	6	94.4 %	101
10	(8;9>	2.8 %	3	97.2 %	104

8.2.6 Task 06.

Table 26: TASK 06				
Analytical Question: How do changes in treatment and E-scores affect patient scores as measured by attributes and features in TRT (Initial patient questinnnaire values)?				
INPUT				
	Antecedent	Succedent		
Stable Part	BASIC, QQQ	Not Used		
Variable Part	TRTM	IMPR_TRT		
OUTPUT				
Number of rules found:	274	Number of verifications:	8496	
Duration (PC dependent):		0h 0m 12s		
Analysis of Rule of Interest (No. 264, ID 242)				
Antecedent: Sc_T(moderate) : (Chg_in_e3(0) -> Chg_in_e3(>0))				
Succedent: (Impr_in_H_An(-) -> Impr_in_H_An(+))				
Condition: (empty)				
State Before: Sc_T(moderate) && Chg_in_e3(0) Impr_in_H_An(-)				
State After: Sc_T(moderate) && Chg_in_e3(>0) Impr_in_H_An(+)				
RESULT GRID (Before and After state)				
	Succedent	¬ Succedent	Succedent	¬ Succedent
Antecedent	9	0	8	8
Support and Confidence				
Support:	Before: .08 (n=107), After: .07 (n=107)			
Confidence:	Before: 1 (9/(9+0)), After: .5 (8/(8+8))			

This rule is similar to many that were generated by the conditions specified. An analysis shows that if the Total Score from the Tinnitus Handicap Inventory shows moderate tinnitus, when the change in emotion is positive as represented by E3 which is the new feature developed from questions related to sleep, then improvement in the annoyance caused by Hyperacusis is realized. From this, a physician may choose to focus on sleep therapy in order to improve the emotions and improve tinnitus. At the very least, this would warrant further study.

8.3 Action Rules and MARDs.

Input to MARDs is quite different from LISp-Miner and 4ft-Miner. The MARDs minimal action rule program requires as input a space delimited file with row 1 being headings and the rest of the rows representing the data to be analyzed. Input consists of a minimum threshold for support and confidence. Stable attributes are indicated by a list of column numbers (0 based) as the first row of the input file representing those attributes and features that are stable. From this initial input, the application stores the headings and data in an array and then builds frequent item sets that meet the support and confidence. After the frequent item sets are generated, action rules are built from a single variable that is indicated as flexible with input including the change state of the variable.

Figure 25 shows the input window for MARDs for an experiment performed with an input file containing features Instrument (INS), ImpinTEL (Improvement in Tinnitus Effect on Life), ChginE1 (Change in E1), ChginE2 (Change in E2), ChginE3 (Change in E3), ChginE4 (Change in E4), and ChginSCT (Change in Score Total). Stable variables were variables 0 to 5, with the Change in Score Total listed as the flexible variable. The change

of interest is the Change in Score Total from a “-“ to a “+” state, representing positive improvement. In all, 107 tuples were entered and input represented the identical input as used in the action rule discovery with LISp-Miner Arc4ft-Miner with the exception of the elimination of all but the categorical variables. LISp-Miner with Arc4ft-Miner has the flexibility to discretize continuous numeric variables and features; this is not present in MARDs.

The input screen for MARDs for the experiment is displayed in Figure 25 below.

```
provide name of file:
actionrulesetoe4.txt
file name entered:actionrulesetoe4.txt;
provide minimum suport rate
.10
provide confidence rate
.4
suport:0.1   confidence:0.4
columns:7   tuples_size749
stable attributes:-1 0
create candidates started
creating 1-itemsets
created:27   1-item sets
2-itemsets started
created:233  2-item sets
3-itemsets started
created:430  3-item sets
4-itemsets started
created:279  4-item sets
5-itemsets started
created:98   5-item sets
6-itemsets started
created:15   6-item sets
7-itemsets started
created:1    7-item sets
k-items candidates created...

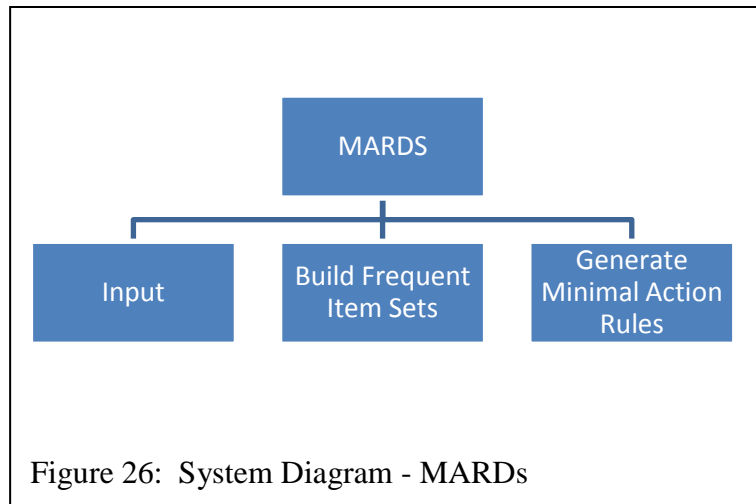
COLUMN NAMES:
1: Ins
2: InmpinTEL
3: ChginE1
4: ChginE2
5: ChginE3
6: ChginE4
7: ChginSCT

In which column do you want to change attribute? <give a number>
3

COLUMN VALUES:
1: -
2: +
Which attribute do you want to change? <give a number>
1
What do you want to change it to? <give a number>
2
RULES CREATED
```

Figure 25: Input Screen for MARDs

A system hierarchy for MARDs is presented in Figure 26.



The input file must be space delimited with no spaces present in headings or data. All input must be discretized prior to processing. Input processing consists of creating arrays (0 based) for column headings and for all rows representing patient visits. The program prompts for minimum support and confidence values. Frequent item sets are built for all combinations, and for this experiment 1,083 action rules were discovered with a minimal support of .10 (107 tuples in the dataset) entered in order to generate rules. The minimum confidence of .4 generated 10 action rules which are presented in Figure 17: MARDs Experiment and Action Rules. The flexible variable was identified as feature E1 with a change from “-“ to “+” indicated.

In order to understand the rules, it is important to understand the reference to the arrays in the program. Rules are presented with array index values and must be interpreted for this experiment based on the particular input file and the order of the values for each column during file input. In other words, if a particular column has a “+” as a value in the

first tuple and a “-“ in the second tuple, the array will list the values as “+” then “-“ for the column 0 and 1 values respectively.

Rule 1: (1,0->1) (4,1->0) (6,1->0) -> (2,1->0) sup:0.102804 conf:1
 Rule 2: (1,0->1) (4,1->0) (6,1) -> (2,1->0) sup:0.102804 conf:1
 Rule 3: (3,0->1) (4,1->0) (6,1->0) -> (2,1->0) sup:0.121495 conf:1
 Rule 4: (3,0->1) (4,1->0) (6,1) -> (2,1->0) sup:0.158879 conf:1
 Rule 5: (0,6) (3,0->1) (4,1->0) (5,1->0) -> (2,1->0) sup:0.102804 conf:1
 Rule 6: (1,0) (3,0->1) (4,1->0) (6,1) -> (2,1->0) sup:0.140187 conf:1
 Rule 7: (1,0->1) (4,1->0) (5,1->0) (6,1) -> (2,1->0) sup:0.102804 conf:1
 Rule 8: (3,0->1) (4,1->0) (5,1->0) (6,1->0) -> (2,1->0) sup:0.11215 conf:1
 Rule 9: (3,0->1) (4,1->0) (5,1->0) (6,1) -> (2,1->0) sup:0.158879 conf:1
 Rule 10: (1,0) (3,0->1) (4,1->0) (5,1->0) (6,1) -> (2,1->0) sup:0.140187 conf:1

Figure 27: MARDs Experiment and Action Rules

Table 27 shows the MARDs input file and the array loading that occurs; this is presented in order to better understand the action rules output from the program.

Table 27: MARDs Input File and Array Loading

Attribute/Feature	0	1	2	3	4	5	6
0 (Ins)	BTE	GH	GHI	GOTE	HAO	Seimans	com
1 (ImpinTEL)	-	+					
2 (ChginE1)	-	+					
3 (ChginE2)	+	-					
4 (ChginE3)	-	+					
5 (ChginE4)	-	+					
6 (ChginSCT)	+	-					

Rules with the highest support and confidence will be discussed. Rule 4 is (3,0->1) (4,1->0) (6,1) -> (2,1->0) with support of .15 and confidence of 1. Rule 4 means if the change in E2 goes from "+" to "-", E3 goes from "+" to "-", and Total Score is "-" then Change in E1 goes from "+" to "-". This shows the relationship of the emotions to the Total Score with the negative emotions in each emotional category being related to a negative Total Score from the Tinnitus Handicap Inventory.

Rule 9 shows support of .16 (rounded) and confidence of 1. Rule 9 shows similar results to Rule 4 and includes emotional feature E4 moving to a negative state with the Total Score reflecting a negative state as well. Additional rules generated from this experiment serve to support the relationship between the emotional features developed as a part of this research and the Total Score from the Tinnitus Handicap Inventory, a measure of patient treatment success. Negative emotions are tied to negative scores/changes on the Tinnitus Handicap Inventory and this discovery is significant.

8.4 A Comparison of Mining Applications.

The primary mining applications utilized in this research were WEKA, LISp Miner with Arc-4ft Miner and MARDs. WEKA was utilized for studies involving classification and association rule discovery with clustered and unclustered data from the original database. With the WEKA study, algorithms for J48, Random Forest, and Multilayer Perceptron were used in the data mining process to evaluate the effectiveness of new features and clustering methods during classification and association rule study.

LISp-Miner further refined the association rules and also allowed action rule discovery with the complex interface provided by the software. New patient data was segmented from the database and used to mine treatment effectiveness based on patient visits, the Total Score from the Tinnitus Handicap Inventory and the new Tinnitus Functional Index. Literally thousands of rules were generated by the studies utilizing LISp-Miner and Arc-4ftMiner. Rules can be quite complex, yet very useful to individuals with expert knowledge in the ontology, such as a physician. The system cost is extensive, and has been documented in [32].

The MARDs system was used on the same dataset developed for LISp-Miner and Arc-4ftMiner with some modifications for the software. MARDs reduces cost by generating minimal rules, as the system discovers rules directly from frequent itemsets generated by the decision system. The limitations of MARDs with respect to discretization of input features and attributes does not limit the usefulness of this important software. Lacking the expert knowledge of the ontology related to the research topic, the data mining researcher can use a tool like MARDs to uncover important minimal action rules thus allowing direction of purpose as a knowledge discovery process is continued.

CHAPTER 9: CONCLUSION AND DISCUSSION

In this dissertation two databases related to Tinnitus Retraining Therapy patients were mined in order to discover knowledge leading to the development of a decision support system for treatment of tinnitus. Tinnitus is a complex problem and the ontology involves domains of neuroscience, human biology, psychology, and audiology.

Preliminary research tasks involved gaining the knowledge necessary to understand the domains in a way necessary to be effective in data mining tasks. After gaining a basic understanding of tinnitus and tinnitus retraining therapy, the next task involved preparing the data for the mining tasks ahead. Data preparation including flattening and clustering the datasets in a manner required in order to effectively use the software applications involving in the research and to uncover knowledge.

Numerous new features were developed based on temporal, numeric and text features in the database. Classical statistical features (standard deviations and averages of hearing tests, primarily loudness discomfort levels) were added to the dataset. Of particular interest were the new temporal features Sound Level Centroid, Sound Level Spread, and Recovery Rate along with categorical features developed from mining text fields in the database. Additionally, four new emotional features were introduced reflecting the relationship between the new Tinnitus Functional Index and the Emotional-Valence plane introduced in Music research. Best classification results on unclustered data were achieved with Sound Level Centroid, Sound Level Spread, and Recovery Rate Features with a

discretized decision variable showing improvement related to the change in Total Score from the Tinnitus Handicap Inventory.

An application of a clustering method for patient visit data was used to introduce homogeneity to the datasets based on time between visits and number of visits. These new clustered datasets had additional features related to the plot of the line of the visit and Total Score from the tinnitus handicap inventory added to the dataset. Coefficients of the polynomial equation representing this line for 3 and 4 visit sets and angles created by the data points (all combinations) were new features added to individual tuples and mined for knowledge and understanding. Angles improved classification for the clustered data.

A new method of learning action rules is proposed as an important part of this study. The MARDs action rule discovery system discovers minimal action rules allowing insight into the relationships that improve treatment in the tinnitus database. MARDs was applied to new patient information and also introduced were several new decision features related to emotions. From the knowledge gained with MARDs, the recommendation is to apply further study to LISp-Miner with Ac-4ft Miner in order to maximize the application of the domain knowledge to the mining process for the knowledge engineer. LISp-Miner shows great promise for uncovering action rules showing the relationship of the emotions to treatment success.

In summary, from the contributions listed the most important of these are the new features that predict treatment success (Sound Level Centroid, Sound Level Spread, Recovery Rate and emotion based features), the link of the emotion based features to the Thayer emotion-valence plane used in music classification, and the system of extracting minimal action rules to facilitate domain knowledge for further and more complete action

rule study. We intend to continue this important work by using the knowledge gained from the extracted action rules to form the basis of a treatment decision support system. This decision support system would apply action rules built on the backbone of this study and the new and improved predictors of tinnitus treatment success related to emotions to input data on the patient during each visit. The recommendation for the patient treatment would be based on the placement of the current patient state to the action rules suggesting treatment patterns to the physician. Information on instruments, emotions, audiological tests, and even medications (after further research) can uncover important relationships and action rules that can predict a pattern of improvement for the patient. In order to build this decision support system, this research should be continued with an expanded dataset.

REFERENCES

- [1] Folmer, R.L. (2002) Long-term reductions in tinnitus severity. *BMC Ear, Nose and Throat Disorders*, p. 1.
- [2] Jastreboff, P.J., Hazell, J.W.P. (2004) *Tinnitus Retraining Therapy: Implementing the Neurophysical Model*, Cambridge University Press, p. 4.
- [3] Jastreboff, P.J., Hazell, J.W.P. (2004) *Tinnitus Retraining Therapy: Implementing the Neurophysical Model*, Cambridge University Press, p. 12-13.
- [4] Henry, J.A., Jastreboff, M.M., Jastreboff, P.J., Schechter, M.A., Fausti, S.A. (2003). Guide to Conducting Tinnitus Retraining Therapy Initial and Follow-Up Interviews. *Journal of Rehabilitation Research and Development*, Vol. 40, No. 2, March/April 2003, p. 159-160.
- [5] Baguley, D.M. (2006) What Progress Have We Made With Tinnitus?, *Acta Oto-Laryngologica*, 126: 5.
- [6] Snow, J.B. Jr. (2006) Strategies of the Tinnitus Research Consortium. *Acta Oto-Laryngologica*, 126: 90.
- [7] Grekow, J., Ras, Z.W. (2010) Emotion Based MIDI Files Retrieval System. *Advances in Music Information Retrieval, Studies in Computational Intelligence*, Vol. 274, Springer, 261-284
- [8] Davis, P.B., Paki, B., Hanley, P.J. (2007) Neuromonics Tinnitus Treatment: Third Clinical Trial. *Ear and Hearing*, Vol. 28, No. 2, p. 242.
- [9] Jastreboff, P.J., Jastreboff, M.M. Tinnitus and Hyperacusis in Ballenger's *Otorhinolaryngology*.
- [10] Henry, J.A., Jastreboff, M.M., Jastreboff, P.J., Schechter, M.A., and Fausti, S. (2003). Guide to Conducting Tinnitus Retraining Therapy Initial and Follow-Up Interviews. *Journal of Rehabilitation Research and Development*, Vol. 40, No. 2, March/April 2003, p. 159-160.
- [11] McCombe, A. (1999). Guidelines for the grading of tinnitus severity. Retrieved from <http://www.otohns.net>.
- [12] Jastreboff, P.J., Jastreboff, M.M. (2005) Tinnitus and Hyperacusis, Chapter 22 in Ballenger's *Otorhinolaryngology Head and Neck Surgery*, 16th ed., p. 470.

- [13] Zhang, X., Thompson, P., Ras, Z.W., Jastreboff, P. Mining tinnitus data based on clustering and new temporal features, in *Learning Structure and Schemas from Documents*, M. Biba, F. Xhafa (Eds.), *Studies in Computational Intelligence*, Vol. 375, Springer, 2011
- [14] Jastreboff, P. (2007, February 13). Interview by plt Thompson [Personal Interview]. Tinnitus study.
- [15] Understanding your hearing test. (2011). Retrieved from <http://www.earinfo.com>.
- [16] Waldon, M.G. (2004) Estimation of Average Stream Velocity, in *J. Hydr. Engrg.*, Volume 130, Issue 11, pp. 1119-1122 (Nov. 2004).
- [17] Zhang, X., Ras, Z.W. (2006). Differentiated Harmonic Feature Analysis on Music Information Retrieval for Instrument Recognition, *Proceedings of IEEE International Conference on Granular Computing (IEEE GrC 2006)*, May 10-12, Atlanta, Georgia, 578-581.
- [18] Povinelli, R.J., Feng, X. (1998) "Temporal Pattern Identification of Time Series Data using Pattern Wavelets and Genetic Algorithms" *Artificial Neural Networks in Engineering*, *Proceedings*, 691-696. data using pattern wavelets and genetic algorithms.
- [19] Powers, D. and X. Yu (1999) *Statistical Methods for Categorical Data Analysis*, Academic Press.
- [20] Meikle, Mary B., Barbara J. Steward, Susan E. Griest, and James A. Henry. Tinnitus Outcomes Assessment in Trends for Amplification, Vol. 12 Number 3, September 2008, p. 223.
- [21] Ras, Z. W. and Dardzinska, A. (2009), Action Rules Discovery Based on Tree Classifiers and Meta-actions, in *Proceedings of the 18th International Symposium on Foundations of Intelligent Systems*, Prague, Czech Republic, 66-75.
- [22] Tsay, L.-S., Ras, Z.W. (2006), Action rules discovery system DEAR3, in *Foundations of Intelligent Systems*, *Proceedings of ISMIS 2006*, Bari, Italy, LNAI, No. 4203, Springer, 483-492
- [23] Simunek M (2003) Academic KDD Project LISp-Miner. In Abraham A et al (eds) *Advances in Soft Computing - Intelligent Systems Design and Applications*, Springer, Berlin Heidelberg New York
- [24] Rauch, J.: Classes of Four Fold Table Quantifiers. In *Principles of Data Mining and Knowledge Discovery*. Red. Zytkow, J – Quafafou, M. Berlin, Springer Verlag 1998, pp. 203–211.
- [25] Rauch, J. (1998) Four-fold Table Calculi and Missing Information. In *JCIS'98 Proceedings*, (Paul P. Wang, editor), Association for Intelligent Machinery, pp. 375-378

- [26] Ras, Z.W., Wieczorkowska, A. (2000), Action-Rules: How to increase profit of a company, in Principles of Data Mining and Knowledge Discovery, Proceedings of PKDD 2000, Lyon, France, LNAI, No. 1910, Springer, 587-592
- [27] Meikle, M. (2008) Trends in Amplification, Tinnitus Outcomes Assessment, p. 224.
- [28] Weka, textbook, 2006 p. 69, 278.
- [29] Ras, Z.W., Dardzinska, A. (2011) From Data to Classification Rules and Actions, International Journal of Intelligent Systems, Wiley, Vol. 26, Issue 6, 2011, 572-590
- [30] Rauch, J., Simunek, M. (2009) Action Rules and the GUHA Method: Preliminary Considerations and Results, in Foundations of Intelligent Systems, LNAI, Vol. 5722, Springer, 76-87
- [31] Nekvapil, V. (2010) Data Mining in the Medical Domain: Using the Ac4ft-Miner Procedure, Lambert Academic Publishing

APPENDIX A: ATTRIBUTES, FEATURES, AND DESCRIPTIONS

Note: decision variables are in yellow in last columns on data worksheet - note "NEW PATIENT Y OR N COLUMN"

Attribute/Feature	Description
Patient ID	Patient ID
Visit Num	Visit Number
Visit Date	Visit Date
Problem	Patient Category: T Tinnitus, H Hyperacusis, M: Misophonia in order of importance
Trtmt Cat Patient	Category of Treatment Chosen by Patient 0-tinnitus minimal problem, 1 tinnitus significant problem, 2- tinnitus significant and hearing loss, 3 tinnitus irrelevant hyperacusis significant, 4 w T - prolonged tinnitus, 4 w H prolonged exacerbation of hyperacusis
Trtmt Cat Dr	Category of Treatment Assigned by Doctor
Miso	Misophonia Y or N (fear of sounds)
miso treat	Treated for Misophonia 1=1, 2=2, 3=1+2
Instrument	Instrument type from visits and contacts (jastreboff says type is most important) type of instruments V - Viennatone, GS - GSI soft, GH - GHI hard, HA - hearing aids, blank - none
D I	Date Instrument Fitted
FU	type of follow up contact, A - audiology and counseling, C - couns, T - telephone, E - E-mail, blank - initial visit
F-1	THI or <i>Neuman Questionnaire</i> scored as 4=yes 2=sometimes 0=no
F-2	Difficult to concentrate? 4 - yes 2 - sometimes 0 - no
E-3	the lower the better
F-4	Difficult to hear people?
C-5	Tinnitus make you angry?
E-6	Tinnitus make you confused?
F-7	Tinnitus make you feel desperate?
C-8	Do you complain a great deal about your tinnitus?
F-9	Trouble falling asleep at night?
E-10	Do you feel like you cannot escape your tinnitus?
C-11	Does tinnitus interfere with your ability to enjoy social activities?
F-12	Tinnitus make you feel frustrated?
E-13	Tinnitus make you feel like you have a terrible disease?
F-14	Tinnitus make it difficult for you to enjoy life?
E-15	Tinnitus interfere with your job or household responsibilities?
F-16	Tinnitus make you often irritable?
E-17	Tinnitus make it difficult for you to read?
C-18	Tinnitus make you upset?
F-19	Tinnitus has caused stress on your relationships with family and friends?
E-20	Difficult to focus attention away from tinnitus and on to other things?
C-21	Do you feel you have no control over your tinnitus?
F-22	Tinnitus make you often feel tired?

E-21	Tinnitus make you feel depressed?
E-22	Tinnitus make you feel anxious?
C-23	Do you feel that you can no longer cope with your tinnitus?
F-24	Does your tinnitus get worse when you are under stress?
E-25	Does your tinnitus make you feel insecure?
Sc F	total score function
Sc E	total score emotion
Sc C	total score catastrophic
Sc T	sum of the above: 0to16 slight severity, 18 to 36 mild, 38 to 56 moderate, 58to76 severe, 78to100catastrophic
LR50	Loudness Discomfort Levels Right and Left Ear Tests
LR1	normal is 90 to110, the higher the better the patient is,
LR2	102 is average or normal,
LR3	81.7 is the average for ppl with decreased sound tolerance
LR4	
LR6	
LR8	
LR12	
L RTP	
LL50	
LL1	
LL2	
LL3	
LL4	
LL6	
LL8	
LL12	
LLTP	
H Sv	<i>Questions from TRT original interview</i>
H An	severity of DST, average over last month, 0 - 10
H EL	DST is decreased sound tolerance
H pr	annoyance of DST average over last month, 0 - 10
HL pr	effect of life of DST, average over last month, 0 - 10
Pr	Hyperacusis as a problem, average over last month, 0 - 10
Aw%T	Hearing Loss as a problem, average over last month, 0 - 10
An%T	program assesment Y - Yes, N - NO, U - unsure
Tch	% of time when aware of Tinnitus over last month
T Sv	% of time when annoyed by Tinnitus over last month
T An	changed? S- same, B- better, W-worse
T EL	severity of tinnitus, average over last month, 0 - 10
	0 is no tinnitus, 10 is as loud as you can imagine
	annoyance of tinnitus, average over last month, 0 -10
	effect of life of tinnitus, average over last month, 0 - 10
	Tinnitus Function Index New Questionnaire
	0 to 10 (10 bad) or %

Q1	% aware
Q2	loud
Q3	in control
Q4	annoyed
Q5	cope
Q6	annoyed
Q7	concentrate
Q8	think clearly
Q9	focus attention
Q10	fall/stay asleep
Q11	as much sleep
Q12	sleeping deeply
Q13	hear clearly
Q14	understand people
Q15	follow conversation
Q16	quite, resting activities
Q17	relax
Q18	peace and quiet
Q19	social activities
Q20	enjoyment of life
Q21	relationships
Q22	work on other tasks
Q23	anxious, worried
Q24	bothered upset
Q25	depressed

APPENDIX B: SAMPLE OF FREQUENT ACTION RULES FROM MARDs

Summary

Total rows in the original set: 107

Total frequent actionrules discovered: 1083

The selected measures: Support=0.1

Rules:

(Ins, GH)
support:0.140187

(Ins, GOTE)
support:0.373832

(Ins, *)
support:0.299065

(InmpinTEL, -)
support:0.598131

(InmpinTEL, -->+)
support:0.401869

(InmpinTEL, +>-)
support:0.401869

(InmpinTEL, +)
support:0.401869

(ChginE1, -)
support:0.626168

(ChginE1, -->+)
support:0.373832

(ChginE1, +>-)
support:0.373832

(ChginE1, +)
support:0.373832

(ChginE2, +)
support:0.514019

(ChginE2, +>-)
support:0.485981

(ChginE2, -->+)
support:0.485981

(ChginE2, -)
support:0.485981

(ChginE3, -)
support:0.616822

. . .

InmpinTEL, -->+) (ChginE1, -) (ChginSCT, -)
support:0.102804

(InmpinTEL, -->+) (ChginE1, -->+) (ChginE2, -->+)
support:0.168224

(InmpinTEL, -->+) (ChginE1, -->+) (ChginE3, -->+)
support:0.168224

(InmpinTEL, -->+) (ChginE1, -->+) (ChginE4, -->+)
support:0.140187

. . .

(InmpinTEL, -) (ChginE1, +) (ChginE4, +) (ChginSCT, -)
support:0.140187

(InmpinTEL, -) (ChginE2, +) (ChginE3, +) (ChginE4, +)
support:0.17757

(InmpinTEL, -) (ChginE2, +) (ChginE3, +) (ChginSCT, -)
support:0.140187

(InmpinTEL, -) (ChginE2, +) (ChginE4, +) (ChginSCT, -)
support:0.149533

(InmpinTEL, -) (ChginE2, +->-) (ChginE3, +->-) (ChginE4, +->-)
support:0.17757

. . .

(InmpinTEL, +->-) (ChginE1, +->-) (ChginE2, +->-) (ChginE3, +->-)
(ChginE4, +->-) (ChginSCT, +->-)
support:0.121495

(InmpinTEL, +->-) (ChginE1, +) (ChginE2, +) (ChginE3, +) (ChginE4,
+) (ChginSCT, +->-)
support:0.121495

(InmpinTEL, +) (ChginE1, +) (ChginE2, +) (ChginE3, +) (ChginE4,
+) (ChginSCT, +)
support:0.121495

(Ins, GOTE) (InmpinTEL, -) (ChginE1, -) (ChginE2, -) (ChginE3, -) (ChginE4,
-) (ChginSCT, -)
support:0.11215